

Alexander Gelbukh (Ed.)

# Computational Linguistics and Intelligent Text Processing

12th International Conference, CICLing 2011  
Tokyo, Japan, February 20-26, 2011  
Proceedings

 Springer

# Table of Contents – Part I

## Lexical Resources

Influence of Treebank Design on Representation of Multiword Expressions . . . . .	1
<i>Eduard Bejček, Pavel Straňák, and Daniel Zeman</i>	
Combining Contextual and Structural Information for Supersense Tagging of Chinese Unknown Words . . . . .	15
<i>Likun Qiu, Yunfang Wu, and Yanqiu Shao</i>	
Identification of Conjunct Verbs in Hindi and Its Effect on Parsing Accuracy . . . . .	29
<i>Rafiya Begum, Karan Jindal, Ashish Jain, Samar Husain, and Dipti Misra Sharma</i>	
Identification of Reduplicated Multiword Expressions Using CRF . . . . .	41
<i>Kishorjit Nongmeikapam, Dhiraj Laishram, Naorem Bikramjit Singh, Ngariyanbam Mayekleima Chanu, and Sivaji Bandyopadhyay</i>	

## Syntax and Parsing

Computational Linguistics and Natural Language Processing (Invited Paper) . . . . .	52
<i>Jun'ichi Tsujii</i>	
An Unsupervised Approach for Linking Automatically Extracted and Manually Crafted LTAGs . . . . .	68
<i>Heshaam Faily and Ali Basirat</i>	
Tamil Dependency Parsing: Results Using Rule Based and Corpus Based Approaches . . . . .	82
<i>Loganathan Ramasamy and Zdeněk Žabokrtský</i>	
Incremental Combinatory Categorical Grammar and Its Derivations . . . . .	96
<i>Ahmed Hefny, Hany Hassan, and Mohamed Bahgat</i>	
Dependency Syntax Analysis Using Grammar Induction and a Lexical Categories Precedence System . . . . .	109
<i>Hiram Calvo, Omar J. Gambino, Alexander Gelbukh, and Kentaro Inui</i>	
Labelwise Margin Maximization for Sequence Labeling . . . . .	121
<i>Wenjun Gao, Xipeng Qiu, and Xuanjing Huang</i>	

Co-related Verb Argument Selectional Preferences (Best Paper Award, First Place) .....	133
<i>Hiram Calvo, Kentaro Inui, and Yuji Matsumoto</i>	

Combining Diverse Word-Alignment Symmetrizations Improves Dependency Tree Projection .....	144
<i>David Mareček</i>	

An Analysis of Tree Topological Features in Classifier-Based Unlexicalized Parsing .....	155
<i>Samuel W.K. Chan, Mickey W.C. Chong, and Lawrence Y.L. Cheung</i>	

## Part of Speech Tagging and Morphology

Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? (Invited Paper) .....	171
<i>Christopher D. Manning</i>	

Ripple Down Rules for Part-of-Speech Tagging .....	190
<i>Dat Quoc Nguyen, Dai Quoc Nguyen, Son Bao Pham, and Dang Duc Pham</i>	

An Efficient Part-of-Speech Tagger for Arabic .....	202
<i>Selçuk Köprü</i>	

An Evaluation of Part of Speech Tagging on Written Second Language Spanish .....	214
<i>M. Pilar Valverde Ibañez</i>	

Onoma: A Linguistically Motivated Conjugation System for Spanish Verbs .....	227
<i>Luz Rello and Eduardo Basterrechea</i>	

## Word Sense Disambiguation

Measuring Similarity of Word Meaning in Context with Lexical Substitutes and Translations (Invited Paper) .....	238
<i>Diana McCarthy</i>	

A Quantitative Evaluation of Global Word Sense Induction .....	253
<i>Marianna Apidianaki and Tim Van de Cruys</i>	

Incorporating Coreference Resolution into Word Sense Disambiguation (Best Student Paper Award) .....	265
<i>Shangfeng Hu and Chengfei Liu</i>	

## Semantics and Discourse

Deep Semantics for Dependency Structures . . . . .	277
<i>Paul Bédaride and Claire Gardent</i>	
Combining Heterogeneous Knowledge Resources for Improved Distributional Semantic Models . . . . .	289
<i>György Szarvas, Torsten Zesch, and Iryna Gurevych</i>	
Improving Text Segmentation with Non-systematic Semantic Relation (Verifiability Award) . . . . .	304
<i>Viet Cuong Nguyen, Le Minh Nguyen, and Akira Shimazu</i>	
Automatic Identification of Cause-Effect Relations in Tamil Using CRFs . . . . .	316
<i>Menaka S., Pattabhi R.K. Rao, and Sobha Lalitha Devi</i>	
Comparing Approaches to Tag Discourse Relations . . . . .	328
<i>Shamima Mithun and Leila Kosseim</i>	
Semi-supervised Discourse Relation Classification with Structural Learning . . . . .	340
<i>Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka</i>	
Integrating Japanese Particles Function and Information Structure . . . . .	353
<i>Akira Ohtani</i>	
Assessing Lexical Alignment in Spontaneous Direction Dialogue Data by Means of a Lexicon Network Model . . . . .	368
<i>Alexander Mehler, Andy Lücking, and Peter Menke</i>	

## Opinion Mining and Sentiment Detection

Towards Well-Gounded Phrase-Level Polarity Analysis . . . . .	380
<i>Robert Remus and Christian Hähnig</i>	
Implicit Feature Identification via Co-occurrence Association Rule Mining . . . . .	393
<i>Zhen Hai, Kuiyu Chang, and Jung-jae Kim</i>	
Construction of Wakamono Kotoba Emotion Dictionary and Its Application . . . . .	405
<i>Kazuyuki Matsumoto and Fuji Ren</i>	
Temporal Analysis of Sentiment Events – A Visual Realization and Tracking . . . . .	417
<i>Dipankar Das, Anup Kumar Kolya, Asif Ekbal, and Sivaji Bandyopadhyay</i>	

## Text Generation

Highly-Inflected Language Generation Using Factored Language Models .....	429
<i>Eder Miranda de Novais, Ivandré Paraboni, and Diogo Takaki Ferreira</i>	
Prenominal Modifier Ordering in Bengali Text Generation .....	439
<i>Sumit Das, Anupam Basu, and Sudeshna Sarkar</i>	
Bootstrapping Multiple-Choice Tests with THE-MENTOR .....	451
<i>Ana Cristina Mendes, Sérgio Curto, and Luísa Coheur</i>	
<b>Author Index</b> .....	463

## Table of Contents – Part II

### Machine Translation and Multilingualism

Ontology Based Interlingua Translation . . . . .	1
<i>Leonardo Lesmo, Alessandro Mazzei, and Daniele P. Radicioni</i>	
Phrasal Equivalence Classes for Generalized Corpus-Based Machine Translation . . . . .	13
<i>Rashmi Gangadharaiah, Ralf D. Brown, and Jaime Carbonell</i>	
A Multi-view Approach for Term Translation Spotting . . . . .	29
<i>Raphaël Rubino and Georges Linarès</i>	
ICE-TEA: In-Context Expansion and Translation of English Abbreviations . . . . .	41
<i>Waleed Ammar, Kareem Darwish, Ali El Kahki, and Khaled Hafez</i>	
Word Segmentation for Dialect Translation . . . . .	55
<i>Michael Paul, Andrew Finch, and Eiichiro Sumita</i>	
TEP: Tehran English-Persian Parallel Corpus . . . . .	68
<i>Mohammad Taher Pilevar, Hesham Faily, and Abdol Hamid Pilevar</i>	
Effective Use of Dependency Structure for Bilingual Lexicon Creation (Best Paper Award, Third Place) . . . . .	80
<i>Daniel Andrade, Takuya Matsuzaki, and Jun'ichi Tsujii</i>	
Online Learning via Dynamic Reranking for Computer Assisted Translation . . . . .	93
<i>Pascual Martínez-Gómez, Germán Sanchis-Trilles, and Francisco Casacuberta</i>	

### Information Extraction and Information Retrieval

Learning Relation Extraction Grammars with Minimal Human Intervention: Strategy, Results, Insights and Plans (Invited Paper) . . . . .	106
<i>Hans Uszkoreit</i>	
Using Graph Based Method to Improve Bootstrapping Relation Extraction . . . . .	127
<i>Haibo Li, Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka</i>	
A Hybrid Approach for the Extraction of Semantic Relations from MEDLINE Abstracts . . . . .	139
<i>Asma Ben Abacha and Pierre Zweigenbaum</i>	

An Active Learning Process for Extraction and Standardisation of Medical Measurements by a Trainable FSA . . . . .	151
<i>Jon Patrick and Mojtaba Sabbagh</i>	
Topic Chains for Understanding a News Corpus . . . . .	163
<i>Dongwoo Kim and Alice Oh</i>	
From Italian Text to TimeML Document via Dependency Parsing . . . . .	177
<i>Livio Robaldo, Tommaso Caselli, Irene Russo, and Matteo Grella</i>	
Self-adjusting Bootstrapping (Best Paper Award, Second Place) . . . . .	188
<i>Shoji Fujiwara and Satoshi Sekine</i>	
Story Link Detection Based on Event Words . . . . .	202
<i>Letian Wang and Fang Li</i>	
Ranking Multilingual Documents Using Minimal Language Dependent Resources . . . . .	212
<i>G.S.K. Santosh, N. Kiran Kumar, and Vasudeva Varma</i>	
Measuring Chinese-English Cross-Lingual Word Similarity with HowNet and Parallel Corpus . . . . .	221
<i>Yunqing Xia, Taotao Zhao, Jianmin Yao, and Peng Jin</i>	

## Text Categorization and Classification

Comparing Manual Text Patterns and Machine Learning for Classification of E-Mails for Automatic Answering by a Government Agency . . . . .	234
<i>Hercules Dalianis, Jonas Sjöbergh, and Eriks Sneiders</i>	
Using Thesaurus to Improve Multiclass Text Classification . . . . .	244
<i>Nooshin Maghsoodi and Mohammad Mehdi Homayounpour</i>	
Adaptable Term Weighting Framework for Text Classification . . . . .	254
<i>Dat Huynh, Dat Tran, Wanli Ma, and Dharmendra Sharma</i>	
Automatic Specialized vs. Non-specialized Sentence Differentiation . . . . .	266
<i>Iria da Cunha, Maria Teresa Cabré, Eric SanJuan, Gerardo Sierra, Juan Manuel Torres-Moreno, and Jorge Vivaldi</i>	
Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features . . . . .	277
<i>B. Thomas Adler, Luca de Alfaro, Santiago M. Mola-Velasco, Paolo Rosso, and Andrew G. West</i>	
Costco: Robust Content and Structure Constrained Clustering of Networked Documents . . . . .	289
<i>Su Yan, Dongwon Lee, and Alex Hai Wang</i>	

## Summarization and Recognizing Textual Entailment

Learning Predicate Insertion Rules for Document Abstracting . . . . .	301
<i>Horacio Saggion</i>	
Multi-topical Discussion Summarization Using Structured Lexical Chains and Cue Words . . . . .	313
<i>Jun Hatori, Akiko Murakami, and Jun'ichi Tsujii</i>	
Multi-document Summarization Using Link Analysis Based on Rhetorical Relations between Sentences . . . . .	328
<i>Nik Adilah Hanin Binti Zahri and Fumiyo Fukumoto</i>	
Co-clustering Sentences and Terms for Multi-document Summarization . . . . .	339
<i>Yunqing Xia, Yonggang Zhang, and Jianmin Yao</i>	
Answer Validation Using Textual Entailment . . . . .	353
<i>Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay</i>	

## Authoring Aid, Error Correction, and Style Analysis

SPIDER: A System for Paraphrasing in Document Editing and Revision — Applicability in Machine Translation Pre-editing . . . . .	365
<i>Anabela Barreiro</i>	
Providing Cross-Lingual Editing Assistance to Wikipedia Editors . . . . .	377
<i>Ching-man Au Yeung, Kevin Duh, and Masaaki Nagata</i>	
Reducing Overdetections in a French Symbolic Grammar Checker by Classification . . . . .	390
<i>Fabrizio Gotti, Philippe Langlais, Guy Lapalme, Simon Charest, and Éric Brunelle</i>	
Performance Evaluation of a Novel Technique for Word Order Errors Correction Applied to Non Native English Speakers' Corpus . . . . .	402
<i>Theologos Athanaselis, Stelios Bakamidis, and Ioannis Dologlou</i>	
Correcting Verb Selection Errors for ESL with the Perceptron . . . . .	411
<i>Xiaohua Liu, Bo Han, and Ming Zhou</i>	
A Posteriori Agreement as a Quality Measure for Readability Prediction Systems . . . . .	424
<i>Philip van Oosten, Véronique Hoste, and Dries Tanghe</i>	
A Method to Measure the Reading Difficulty of Japanese Words . . . . .	436
<i>Keiji Yasuda, Andrew Finch, and Eiichiro Sumita</i>	

Informality Judgment at Sentence Level and Experiments with Formality Score . . . . .	446
<i>Shibamouli Lahiri, Prasenjit Mitra, and Xiaofei Lu</i>	

## Speech Recognition and Generation

Combining Word and Phonetic-Code Representations for Spoken Document Retrieval . . . . .	458
<i>Alejandro Reyes-Barragán, Manuel Montes-y-Gómez, and Luis Villaseñor-Pineda</i>	

Automatic Rule Extraction for Modeling Pronunciation Variation . . . . .	467
<i>Zeeshan Ahmed and Julie Carson-Berndsen</i>	

Predicting Word Pronunciation in Japanese . . . . .	477
<i>Jun Hatori and Hisami Suzuki</i>	

A Minimum Cluster-Based Trigram Statistical Model for Thai Syllabification . . . . .	493
<i>Chonlasith Jucksriporn and Ohm Sornil</i>	

Automatic Generation of a Pronunciation Dictionary with Rich Variation Coverage Using SMT Methods . . . . .	506
<i>Panagiota Karanasou and Lori Lamel</i>	

<b>Author Index</b> . . . . .	519
-------------------------------	-----

# Influence of Treebank Design on Representation of Multiword Expressions

Eduard Bejček, Pavel Straňák, and Daniel Zeman

Charles University in Prague, Institute of Formal and Applied Linguistics  
{bejcek, stranak, zeman}@ufal.mff.cuni.cz

**Abstract.** Multiword Expressions (MWEs) are important linguistic units that require special treatment in many NLP applications. It is thus desirable to be able to recognize them automatically. Semantically annotated corpora should mark MWEs in a clear way that facilitates development of automatic recognition tools. In the present paper we discuss various corpus design decisions from this perspective. We propose guidelines that should lead to MWE-friendly annotation and evaluate them on numerous sentence examples. Our experience of identifying MWEs in the Prague Dependency Treebank provides the base for the discussion and examples from other languages are added whenever appropriate.

## 1 Motivation

Grammatical theories have been thriving recently in computational linguistics. They describe phenomena of natural language in increasing detail with the purpose of creating a description that analyses and/or generates language as natural as possible.

Several treebanks have been developed during the past decade, new ones are still being created and the old ones are being enriched with additional annotations. A corpus is often designed and developed with the vision of further, deeper annotation, with the aim to add semantic information in future. Multiword expressions (MWEs; such as idioms, phrasemes, multiword named entities) are an important part of most natural languages. Usually they form a significant portion of vocabulary, particularly in special domains where terminology is in play, but not only there.

Although some grammatical theories have accounted for MWEs decades ago (see e.g. [1]), in treebanks, multiword expressions are one of the least developed phenomena. Recently, however, their processing started to attract attention, as they are proving to be important for information extraction, machine translation and other crucial tasks of NLP [2]. Therefore they should be an integral part of any serious semantic annotation.

In this paper, we discuss some decisions of a treebank design that have direct influence on representation of MWEs. A good treebank design can contribute to both more natural and more useful representation of MWEs, or even enable to capture certain rare forms of MWEs. We will also discuss the decisions that make the representation of MWEs harder or inefficient (see Section 3).

# Combining Contextual and Structural Information for Supersense Tagging of Chinese Unknown Words

Likun Qiu<sup>1</sup>, Yunfang Wu<sup>1</sup>, and Yanqiu Shao<sup>2</sup>

<sup>1</sup> Key Laboratory of Computational Linguistics (Peking University),  
Ministry of Education 100871 Beijing, China

<sup>2</sup> Institute of Artificial Intelligence, Beijing City University,  
100083 Beijing, China

{qiulikun,wuyf}@pku.edu.cn, yqshao@bcu.edu.cn

**Abstract.** Supersense tagging classifies unknown words into semantic categories defined by lexicographers and inserts them into a thesaurus. Previous studies on supersense tagging show that context-based methods perform well for English unknown words while structure-based methods perform well for Chinese unknown words. The challenge before us is how to successfully combine contextual and structural information together for supersense tagging of Chinese unknown words. We propose a simple yet effective approach to address the challenge. In this approach, contextual information is used for measuring contextual similarity between words while structural information is used to filter candidate synonyms and adjusting contextual similarity score. Experiment results show that the proposed approach outperforms the state-of-art context-based method and structure-based method.

**Keywords:** Supersense Tagging, Contextual Information, Structural Information, Chinese Unknown Words.

## 1 Introduction

Lexical-semantic resources such as *WordNet* [1] have influenced NLP research significantly. These resources have been successfully applied in a wide range of research [2, 3, 4, 5]. However, to keep up with the pace of language evolution, lexicographers should update the resources by hand, which is time-consuming and labor-intensive. To reduce human effort, a technology called supersense tagging [6] is presented to help lexicographers classify unknown words and insert them into an existing resource. Here, *supersense* refers to a semantic class to which the unknown word belongs, e.g., “tool”, “organization”, “person”, etc. A similar name of that task is semantic classification [7].

To address the problem of supersense tagging, two kinds of information might be utilized, i.e., contextual information and structural information.

Pilot studies on English unknown words mainly used contextual information [6, 8]. The methods of these studies are based on the *distributional hypothesis*, which means that words having similar meaning usually appear in similar context. The experiments of these studies show the effectiveness of contextual information for English.

# Identification of Conjunct Verbs in Hindi and Its Effect on Parsing Accuracy

Rafiya Begum, Karan Jindal, Ashish Jain,  
Samar Husain, and Dipti Misra Sharma

Language Technologies Research Centre, IIIT-Hyderabad, India  
rafiyabegum@gmail.com,  
{karan\_jindal, ashishjain}@students.iiit.ac.in,  
{samar, dipti}@mail.iiit.ac.in

**Abstract.** This paper introduces a work on identification of conjunct verbs in Hindi. The paper will first focus on investigating which noun-verb combination makes a conjunct verb in Hindi using a set of linguistic diagnostics. We will then see which of these diagnostics can be used as features in a MaxEnt based automatic identification tool. Finally we will use this tool to incorporate certain features in a graph based dependency parser and show an improvement over previous best Hindi parsing accuracy.

**Keywords:** Conjunct verbs, Diagnostics, Automatic Identification, Parsing, Light verb.

## 1 Introduction

There are certain verbs that need other words in the sentence to represent an activity or a state of being. Such verbs along with the other words, required for completion of meaning, are together called *Complex Predicates* (CP). CP exist in great numbers in South Asian languages [1], [2], [3]. A CP is generally made via the combination of nouns, adjectives and verbs with other verbs. The verb in the CP is referred as light verb and the element that the light verb combines to form a CP is referred as host [4].

[5] says that in Hindi/Urdu, the light verb is taken as a contributing ‘semantic structure’ which determines syntactic information such as case marking whereas *host* contributes the ‘semantic substance’, i.e. most of the meaning the complex predicate has. [6] has talked about four types of complex predicates: (a) In Syntactic Complex Predicates the formation takes place in the syntax. (b) In Morphological Complex Predicates, a piece of morphology is used to modify the primary event predication. Well known example is morphological causatives. (c) Light Verbs cross linguistically do not always form a uniform syntactic category. They are not always associated with a uniform semantics, but they always muck around with the primary event predication. (d) In Semantics, complex predicates represent the decomposition of event structure.

In CPs, ‘Noun/Adjective+Verb’ combinations are called conjunct verbs and ‘Verb+Verb’ combinations are called compound verbs. In this paper, we are focusing

# Identification of Reduplicated Multiword Expressions Using CRF

Kishorjit Nongmeikapam<sup>1</sup>, Dhiraj Laishram<sup>1</sup>, Naorem Bikramjit Singh<sup>1</sup>,  
Ngariyanbam Mayekleima Chanu<sup>2</sup>, and Sivaji Bandyopadhyay<sup>3</sup>

<sup>1</sup> Dept. of Computer Sc. & Engg., Manipur Institute of Technology,  
Manipur University, Imphal, India

<sup>2</sup> Dept. of Education Technology, Kanan Devi Memorial  
College of Education, Imphal, India

<sup>3</sup> Dept. of Computer Sc. & Engg., Jadavpur University,  
Jadavpur, Kolkata, India

{kishorjit.nongmeikapa,dhirajlaishram,  
naorembikramjit10,mayekleima.ng}@gmail.com,  
sivaji\_cse\_ju@yahoo.com

**Abstract.** This paper deals with the identification of Reduplicated Multiword Expressions (RMWEs) which is important for any natural language applications like Machine Translation, Information Retrieval etc. In the present task, reduplicated MWEs have been identified in Manipuri language texts using CRF tool. Manipuri is highly agglutinative in nature and reduplication is quite high in this language. The important features selected for running the CRF tool include stem words, number of suffixes, number of prefixes, prefixes in the word, suffixes in the word, Part Of Speech (POS) of the surrounding words, surrounding stem words, length of the word, word frequency and digit feature. Experimental results show the effectiveness of the proposed approach with the overall average Recall, Precision and F-Score values of 92.91%, 91.90% and 92.40% respectively.

**Keywords:** Multiword Expressions (MWE), Reduplicated MWE, Conditional Random Field (CRF), Manipuri.

## 1 Introduction

Manipuri (or Meiteilon) belongs to the Tibeto-Burman family of languages mainly spoken in Manipur, a state in the North East India. Manipuri is a schedule language of Indian constitution. This language is also spoken in some parts of the other countries like Myanmar and Bangladesh. Manipuri is highly agglutinative in nature, monosyllabic, influenced and enriched by the Indo-Aryan languages of Sanskrit origin and English. The affixes play the most important role in the structure of the language. The majority of the roots found in the language are bound and the affixes are the determining factor of the class of the words in the language. Manipuri uses two scripts; the first one is purely of its own origin (Meitei Mayek) while another one is a borrowed Bengali script. In the present task, the processing has been done on the Bengali script.

# Computational Linguistics and Natural Language Processing

Jun'ichi Tsujii

Department of Computer Science, University of Tokyo  
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan  
School of Computer Science, University of Manchester  
National Centre for Text Mining (NaCTeM)  
Manchester Interdisciplinary Biocentre, 131 Princess Street,  
Manchester M1 7DN, UK  
tsujii@is.s.u-tokyo.ac.jp

**Abstract.** Researches in Computational Linguistics (CL) and Natural Language Processing (NLP) have been increasingly dissociated from each other. Empirical techniques in NLP show good performances in some tasks when large amount of data (with annotation) are available. However, in order for these techniques to be adapted easily to new text types or domains, or for similar techniques to be applied to more complex tasks such as text entailment than POS taggers, parsers, etc., rational understanding of language is required. Engineering techniques have to be underpinned by scientific understanding. In this paper, taking grammar in CL and parsing in NLP as an example, we will discuss how to re-integrate these two research disciplines. Research results of our group on parsing are presented to show how grammar in CL is used as the backbone of a parser.

## 1 Introduction

The two terms, Computational Linguistics (CL) and Natural Language Processing (NLP), have often been used interchangeably. However, these two terms represent two different streams of research which emphasize different aspects of our field. For example, while research on grammar and its formalisms in CL and research on parsing in NLP are closely related, their objectives are quite different. On one hand, researchers in CL have focused on revealing how surface strings of words systematically correspond to their meanings (in a compositional way) and have been interested in developing formalisms by which the correspondences are described. On the other hand, those in NLP are interested in more practical engineering issues involved in processing natural languages by computer, such as efficient algorithms for a program (parser) which computes the structure and/or the meaning of a given sentence.

While some of parsers used in NLP are based on a grammar in CL, in order for them to be practical, they should not only be efficient and robust but also be able to choose the most plausible interpretation of a sentence among many possible

# An Unsupervised Approach for Linking Automatically Extracted and Manually Crafted LTAGs

Heshaam Faili and Ali Basirat

Department of ECE, University of Tehran, Tehran, Iran  
hfaili@ut.ac.ir

Department of Computer Engineering, Islamic Azad University,  
Science and Research, Branch of Tehran, Iran  
a.basirat@srbiau.ac.ir

**Abstract.** Though the lack of semantic representation of automatically extracted LTAGs is an obstacle in using these formalism, due to the advent of some powerful statistical parsers that were trained on them, these grammars have been taken into consideration more than before. Against of this grammatical class, there are some widely usage manually crafted LTAGs that are enriched with semantic representation but suffer from the lack of efficient parsers. The available representation of latter grammars beside the statistical capabilities of former encouraged us in constructing a link between them.

Here, by focusing on the automatically extracted LTAG used by MICA [4] and the manually crafted English LTAG namely XTAG grammar [32], a statistical approach based on HMM is proposed that maps each sequence of former elementary trees onto a sequence of later elementary trees. To avoid of converging the HMM training algorithm in a local optimum state, an EM-based learning process for initializing the HMM parameters were proposed too. Experimental results show that the mapping method can provide a satisfactory way to cover the deficiencies arises in one grammar by the available capabilities of the other.

**Keywords:** Supertagging, HMM Initialization, XTAG Derivation Tree, Semantic Representation, Grammar Mapping, Automatically Extracted Tree Adjoining Grammar, MICA.

## 1 Introduction

Tree Adjoining Grammar (TAG) introduced by Joshi [17] as a Mildly Context Sensitive Grammar is supposed to be powerful enough to model the natural languages [18]. In Lexicalized case (Lexicalized Tree Adjoining Grammar, LTAG), any elementary tree of a TAG could be considered as a complex description of the lexical item that provides a domain of locality, which specifies syntactic and semantic dependencies [2]. Lexicalization, Extended Domain of Locality (EDL) and Factoring of Recursion from the Domain of dependencies (FRD) are three important keys of using this formalism in NLP applications and theorems.

# Tamil Dependency Parsing: Results Using Rule Based and Corpus Based Approaches

Loganathan Ramasamy and Zdeněk Žabokrtský

Institute of Formal and Applied Linguistics (ÚFAL)  
Faculty of Mathematics and Physics, Charles University in Prague  
{ramasamy,zabokrtsky}@ufal.mff.cuni.cz

**Abstract.** Very few attempts have been reported in the literature on dependency parsing for Tamil. In this paper, we report results obtained for Tamil dependency parsing with rule-based and corpus-based approaches. We designed annotation scheme partially based on Prague Dependency Treebank (PDT) and manually annotated Tamil data (about 3000 words) with dependency relations. For corpus-based approach, we used two well known parsers MaltParser and MSTParser, and for the rule-based approach, we implemented series of linguistic rules (for resolving coordination, complementation, predicate identification and so on) to build dependency structure for Tamil sentences. Our initial results show that, both rule-based and corpus-based approaches achieved the accuracy of more than 74% for the unlabeled task and more than 65% for the labeled tasks. Rule-based parsing accuracy dropped considerably when the input was tagged automatically.

**Keywords:** Tamil, Dependency parsing, Syntax, Clause boundaries.

## 1 Introduction

The most important thing in Natural Language Processing (NLP) research is data, importantly the data annotated with linguistic descriptions. Much of the success in NLP in the present decade can be attributed to data driven approaches to linguistic challenges, which discover rules from data as opposed to traditional rule based paradigms. The availability of annotated data such as Penn Treebank [1] and parallel corpora such as Europarl [2] had spurred the application of statistical techniques [4], [5], [3] to various tasks such as Part Of Speech (POS) tagging, syntactic parsing and Machine Translation (MT) and so on. They produced state of the art results compared to their rule based counterparts. Unfortunately, only English and very few other languages have the privilege of having such rich annotated data due to various factors.

In this paper, we take up the case of dependency parsing task for Tamil language for which no annotated data is available. We report our initial results of applying rule based and corpus based techniques to this task. For rule-based approach, the rules (such as rules for coordination and main predicate identification) have been crafted after studying the Tamil data in general. The rules often rely on morphological cues to identify governing or dependent nodes. For corpus-based approach, we

# Incremental Combinatory Categorical Grammar and Its Derivations<sup>\*</sup>

Ahmed Hefny<sup>1</sup>, Hany Hassan<sup>2</sup>, and Mohamed Bahgat<sup>3</sup>

<sup>1</sup> Computer Engineering Department, Faculty of Engineering, Cairo University, Giza, Egypt  
ahefny@eng.cu.edu.eg

<sup>2</sup> Microsoft Research, Redmond, WA 98052, USA  
hanyh@microsoft.com

<sup>3</sup> IBM Cairo Technology Development Center, P.O. 166 El-Ahram, Giza, Egypt  
mbahgat@eg.ibm.com

**Abstract.** Incremental parsing is appealing for applications such as speech recognition and machine translation due to its inherent efficiency as well as being a natural match for the language models commonly used in such systems. In this paper we introduce an Incremental Combinatory Categorical Grammar (ICCG) that extends the standard CCG grammar to enable fully incremental left-to-right parsing. Furthermore, we introduce a novel dynamic programming algorithm to convert CCGbank normal form derivations to incremental left-to-right derivations and show that our incremental CCG derivations can recover the unlabeled predicate-argument dependency structures with more than 96% F-measure. The introduced CCG incremental derivations can be used to train an incremental CCG parser.

## 1 Introduction

An incremental parser is able to process an input sentence left-to-right, word-by-word, and build for each prefix of the input sentence a partial parse that is a sub-graph of the partial parse that it builds for a longer prefix. Besides being cognitively plausible, an incremental parser is more appealing for applications since its time and space complexities are close to linear in input length. It should, for example, constitute a natural match for the word-by-word decoding and pruning schemes used within statistical machine translation and speech recognition.

Combinatory Categorical Grammar (CCG) [10] is a lexicalized grammar formalism that has very strong potential for incremental parsing, mainly due to its ability to represent an arbitrary subsequence of a valid sentence by a single category even if they do not form a complete phrase. CCG [10] extends the categorial grammar by adding new combinatory rules such as type raising and composition. Not only do these extensions increase the grammar coverage and ability to recover long-range dependencies but also they allow incremental parsing. However, there are still difficulties in handling some linguistic constituents such as back modifiers (e.g. adverbs) in an efficient and deterministic, yet incremental manner. Although standard CCG rules may be able to handle

---

<sup>\*</sup> This work was conducted while the first two authors were at IBM Cairo Technology Development Center.

# Dependency Syntax Analysis Using Grammar Induction and a Lexical Categories Precedence System\*

Hiram Calvo<sup>1,2</sup>, Omar J. Gambino<sup>1</sup>, Alexander Gelbukh<sup>1,3</sup>, and Kentaro Inui<sup>2</sup>

<sup>1</sup> Center for Computing Research, IPN,

Av. J. D. Bátiz e/ M.O. de Mendizábal, México, D.F., 07738. México

<sup>2</sup> Computational Linguistics, Nara Institute of Science and Technology,  
Takayama, Ikoma, Nara 630-0192, Japan

<sup>3</sup> Faculty of Law, Waseda University,

Nishi-Waseda 1-6-1, Shinjuku-ku, Tokyo 169-8050, Japan

hcalvo@cic.ipn.mx, omarjg82@gmail.com,

www.gelbukh.com, inui@is.naist.jp.

http://www.diluct.com

**Abstract.** The unsupervised approach for syntactic analysis tries to discover the structure of the text using only raw text. In this paper we explore this approach using Grammar Inference Algorithms. Despite of still having room for improvement, our approach tries to minimize the effect of the current limitations of some grammar inductors by adding morphological information before the grammar induction process, and a novel system for converting a shallow parse to dependencies, which reconstructs information about inductor's undiscovered heads by means of a lexical categories precedence system. The performance of our parser, which needs no syntactic tagged resources or rules, trained with a small corpus, is 10% below to that of commercial semi-supervised dependency analyzers for Spanish, and comparable to the state of the art for English.

## 1 Introduction

There are mainly two approaches for creating syntactic dependencies analyzers: supervised and unsupervised. The main goal of the first approach is to attain the best possible performance for a single language. For this purpose, a great collection of resources is collected (manually annotated corpora with part of speech annotation, and syntactic and structure tags) which requires a great effort and years to be collected. For this approach the state of the art is around of 85% of syntactic annotation of full sentences in several languages (Sabine and Marsi, 2006), getting over 90% for English. On the other hand, the unsupervised approach tries to discover the structure of the text using only raw text. This would allow creating a dependency

---

\* We thank the support of SNI, SIP-IPN, COFAA-IPN, and PIFI-IPN, CONACYT; and the Japanese Government; the first author is a JSPS fellow. The third author is a Visiting Scholar at Waseda University and specifically acknowledges support of SIP-20100773 grant, CONACYT 50206-H grant, and CONACYT scholarship for Sabbatical stay 2010.

# Labelwise Margin Maximization for Sequence Labeling

Wenjun Gao, Xipeng Qiu, and Xuanjing Huang

School of Computer Science, Fudan University, China  
{082024008, xpqiu, xjhuang}@fudan.edu.cn

**Abstract.** In sequence labeling problems, the objective functions of most learning algorithms are usually inconsistent with evaluation measures, such as Hamming loss. In this paper, we propose an online learning algorithm that addresses the problem of labelwise margin maximization for sequence labeling. We decompose the sequence margin to per-label margins and maximize these per-label margins individually, which can result to minimize the Hamming loss of sequence. We compare our algorithm with three state-of-art methods on three tasks, and the experimental results show our algorithm outperforms the others.

## 1 Introduction

In recent years, the sequence labeling problems have obtained much attention especially in the machine learning, computational biology and natural language processing communities such as part-of-speech tagging [16], chunking [17], named entity recognition [18] and Chinese word segmentation [26,15]. The goal of sequence labeling is to assign labels to all elements of a sequence. Due to the exponential size of the output space, sequence labeling problems tend to be more challenging than the conventional classification problems.

Recently, many algorithms have been applied for sequence labeling and the progress has been encouraging. These algorithms usually have the different objective functions. For example, average perceptron algorithm [1] aims to minimize the 0-1 loss of sequence. Maximum entropy Markov models (MEMM) [13] and conditional random fields (CRF) [11] aims to maximize the conditional likelihood. SVM<sup>struct</sup> [23] and maximum margin Markov networks (M3N) [22] aims to maximize the margin or minimize hinge loss [5].

However, most of these objective functions often calculate the conditional probability or margin on the whole sequence, which are usually inconsistent with conventional evaluation measures of sequence labeling, such as Hamming loss. The probability and margin cannot response the Hamming loss directly.

	an	exciting	moment						
Sentence: X	一	个	激	动	人	心	的	时	刻
Correct Label: $L_c$	B	E	B	M	M	E	S	B	E
Wrong Label: $L_w$	B	E	B	E	B	E	S	B	E
Margin: $M_1=9$	1	1	1	1	1	1	1	1	1
Margin: $M_2=12$	1	1	4	-1	-1	4	1	1	1

**Fig. 1.** Per-label decomposition of margin for Chinese word segmentation. “B”, “M”, “E” and “S”, which represent the beginning, middle, end or single character of a word respectively.

# Co-related Verb Argument Selectional Preferences\*

Hiram Calvo<sup>1</sup>, Kentaro Inui<sup>2</sup>, and Yuji Matsumoto<sup>3</sup>

<sup>1,3</sup> Computational Linguistics, Nara Institute of Science and Technology,  
Takayama, Ikoma, Nara 630-0192, Japan

<sup>2</sup> Communication Science Lab., Tohoku University  
Aoba, Sendai, 980-8579, Japan  
{calvo,matsu}@is.naist.jp, inui@ecei.tohoku.ac.jp

**Abstract.** Learning Selectional Preferences has been approached as a verb and argument problem, or at most as a tri-nary relationship between subject, verb and object. The correlation of all arguments in a sentence, however, has not been extensively studied for sentence plausibility measuring because of the increased number of potential combinations and data sparseness. We propose a unified model for machine learning using SVM (Support Vector Machines) with features based on topic-projected words from a PLSI (Probabilistic Latent Semantic Indexing) Model and PMI (Pointwise Mutual Information) as co-occurrence features, and WordNet top concept projected words as semantic classes. We perform tests using a pseudo-disambiguation task. We found that considering all arguments in a sentence improves the correct identification of plausible sentences with an increase of 10% in recall among other things.

## 1 Introduction

A sentence can be regarded as a verb with multiple arguments. The plausibility of each argument depends not only on the verb, but also on other arguments. Measuring the plausibility of verb arguments is required for several tasks such as Semantic Role Labeling, since grouping verb arguments and measuring their plausibility increases performance, as shown by Merlo and Van Der Plas (2009) and Deschacht and Moens (2009). Metaphora recognition requires this information too, since we are able to know common usages of arguments, and an uncommon usage would suggest its presence, or a coherence mistake (*v. gr. to drink the moon in a glass*). Malapropism detection can use the measure of the plausibility of an argument to determine misuses of words (Bolshakov, 2005) as in hysteric *center*, instead of *historic center*; *density has brought me to you*; *It looks like a tattoo subject*; and *Why you say that with ironing?* Anaphora resolution consists on finding referenced objects, thus, requiring among other things, to have information about the plausibility of arguments at hand, *i.e.*, what kind of fillers is more likely to satisfy the sentence's constraints, such as in: *The boy plays with it there, It eats grass, I drank it in a glass.*

This problem can be seen as collecting a large database of semantic frames with detailed categories and examples that fit these categories. For this purpose, recent

---

\* This research is supported by SNI, SIP-IPN, COFAA-IPN, and PIFI-IPN, CONACYT; and the Japanese Government (JSPS).

# Combining Diverse Word-Alignment Symmetrizations Improves Dependency Tree Projection

David Mareček

Charles University in Prague,  
Institute of Formal and Applied Linguistics  
marecek@ufal.mff.cuni.cz

**Abstract.** For many languages, we are not able to train any supervised parser, because there are no manually annotated data available. This problem can be solved by using a parallel corpus with English, parsing the English side, projecting the dependencies through word-alignment connections, and training a parser on the projected trees. In this paper, we introduce a simple algorithm using a combination of various word-alignment symmetrizations. We prove that our method outperforms previous work, even though it uses McDonald’s maximum-spanning-tree parser as it is, without any “unsupervised” modifications.

## 1 Introduction

Syntactic parsing is one of the basic tasks in natural language processing. The best parsers learn grammar from manually annotated treebanks and for all there holds the rule that increasing amount of training data improves their performance. However, there are many languages for which a very few linguistic resources exist. Developing a new treebank is quite expensive and for some rarer languages it is even impossible to find linguists for the annotations.

In recent years, numerous works have been devoted to developing parsers that would not need much annotated data. One way is the totally unsupervised parsing (e.g. the Klein and Manning’s inside-outside method [1]), which infers the dependencies from raw texts only. But the performance of such parsers is quite low so far. This changes when we append a few annotated sentences to the raw texts. Koo proved in [2] that this causes a great improvement.

Hwa came up with an idea [3] to use a parallel corpus. For many languages there exist some form of parallel texts, very often with English or other resource-rich languages being the coupled. The idea is to make a word-alignment, parse the English side of the corpus, then project the dependencies from English to the other language using the alignment, and, finally, train a parser on the resulting trees or tree fragments. Several similar works ([4], [5], [6], [7]) came after and made many improvements on this process. Ganchev [6] and Smith and Eisner [5] combine this method with unsupervised training.

# An Analysis of Tree Topological Features in Classifier-Based Unlexicalized Parsing

Samuel W.K. Chan<sup>1</sup>, Mickey W.C. Chong<sup>1</sup>, and Lawrence Y.L. Cheung<sup>2</sup>

<sup>1</sup> Dept. of Decision Sciences

<sup>2</sup> Dept. of Linguistics & Modern Languages,  
Chinese University of Hong Kong,  
Shatin, Hong Kong SAR

{swkchan, mickey\_chong, yllcheung}@cuhk.edu.hk

**Abstract.** A novel set of “tree topological features” (TTFs) is investigated for improving a classifier-based unlexicalized parser. The features capture the location and shape of subtrees in the treebank. Four main categories of TTFs are proposed and compared. Experimental results showed that each of the four categories independently improved the parsing accuracy significantly over the baseline model. When combined using the ensemble technique, the best unlexicalized parser achieves 84% accuracy without any extra language resources, and matches the performance of early lexicalized parsers. Linguistically, TTFs approximate linguistic notions such as grammatical weight, branching property and structural parallelism. This is illustrated by studying how the features capture structural parallelism in processing coordinate structures.

**Keywords:** parsing, unlexicalized model, topological features, machine learning.

## 1 Introduction

Advances in parsing have been made on two major fronts, namely, learning models and algorithms, and parsing features. In addition to improving probabilistic modelling and classifier-based methods, new parsing features have been developed in the last two decades, for example, the propagation of lexical head feature (Magerman, 1995; Collins, 2003) and semantic features. This paper explores a novel set of features, collectively called “tree topological features” (TTFs). TTFs can be easily computed with no extra language resources and be integrated into parsing models. They also deliver significant improvement over the baseline model.

This study is motivated by the fact that mainstream parsers seldom consider the shape of subtrees dominated by these nodes and rely primarily on matching POS/syntactic tags. As a result, an NP with a complicated structure is treated the same as an NP that dominates only one word. However, linguists working in syntactic processing have long observed that the size and shape of subtree also affect word ordering and parse tree building. For example, (i) branching property, (ii) heaviness of

# Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?

Christopher D. Manning

Departments of Linguistics and Computer Science,  
Stanford University,  
353 Serra Mall, Stanford CA 94305-9010  
manning@stanford.edu

**Abstract.** I examine what would be necessary to move part-of-speech tagging performance from its current level of about 97.3% token accuracy (56% sentence accuracy) to close to 100% accuracy. I suggest that it must still be possible to greatly increase tagging performance and examine some useful improvements that have recently been made to the Stanford Part-of-Speech Tagger. However, an error analysis of some of the remaining errors suggests that there is limited further mileage to be had either from better machine learning or better features in a discriminative sequence classifier. The prospects for further gains from semi-supervised learning also seem quite limited. Rather, I suggest and begin to demonstrate that the largest opportunity for further progress comes from improving the taxonomic basis of the linguistic resources from which taggers are trained. That is, from improved descriptive linguistics. However, I conclude by suggesting that there are also limits to this process. The status of some words may not be able to be adequately captured by assigning them to one of a small number of categories. While conventions can be used in such cases to improve tagging consistency, they lack a strong linguistic basis.

## 1 Isn't Part-of-Speech Tagging a Solved Task?

At first glance, current part-of-speech taggers work rapidly and reliably, with per-token accuracies of slightly over 97% [1–4]. Looked at more carefully, the story is not quite so rosy. This evaluation measure is easy both because it is measured per-token and because you get points for every punctuation mark and other tokens that are not ambiguous. It is perhaps more realistic to look at the rate of getting whole sentences right, since a single bad mistake in a sentence can greatly throw off the usefulness of a tagger to downstream tasks such as dependency parsing. Current good taggers have sentence accuracies around 55–57%, which is a much more modest score. Accuracies also drop markedly when there are differences in topic, epoch, or writing style between the training and operational data.

Still, the perception has been that same-epoch-and-domain part-of-speech tagging is a solved problem, and its accuracy cannot really be pushed higher. I

# Ripple Down Rules for Part-of-Speech Tagging

Dat Quoc Nguyen<sup>1</sup>, Dai Quoc Nguyen<sup>1</sup>,  
Son Bao Pham<sup>1,2</sup>, and Dang Duc Pham<sup>1</sup>

<sup>1</sup> Human Machine Interaction Laboratory,  
Faculty of Information Technology,  
University of Engineering and Technology,  
Vietnam National University, Hanoi  
{datnq,dainq,sonpb,dangpd}@vnu.edu.vn

<sup>2</sup> Information Technology Institute,  
Vietnam National University, Hanoi

**Abstract.** This paper presents a new approach to learn a rule based system for the task of part of speech tagging. Our approach is based on an incremental knowledge acquisition methodology where rules are stored in an exception-structure and new rules are only added to correct errors of existing rules; thus allowing systematic control of interaction between rules. Experimental results of our approach on English show that we achieve in the best accuracy published to date: 97.095% on the Penn Treebank corpus. We also obtain the best performance for Vietnamese VietTreeBank corpus.

## 1 Introduction

Part-of-speech (POS) tagging is one of the most important tasks in Natural Language Processing, which assigns a tag representing its lexical category to each word in a text. After the text is tagged or annotated, it can be used in many applications such as: machine translation, information retrieval etc. A number of approaches for this task have been proposed that achieved state-of-the-art results including: Hidden Markov Model-based approaches [1], Maximum Entropy Model-based approaches [2] [3] [4], Support Vector Machine algorithm-based approaches [5], Perceptron learning algorithms [2][6]. All of these approaches are complex statistics-based approaches while the obtained results are progressing to the limit. The combination utilizing the advantages of simple rule-based systems [7] can surpass the limit. However, it is difficult to control the interaction among a large number of rules.

Brill [7] proposed a method to automatically learn transformation rules for the POS tagging problem. In Brill's learning, the selected rule with the highest score is learned on the context that is generated by all preceding rules. In additions, there are interactions between rules with only front-back order, which means an applied back rule will change the results of all the front rules in the whole text. Hepple [8] presented an approach with two assumptions for disabling interactions between rules to reduce the training time while sacrificing a small fall of accuracy.

# An Efficient Part-of-Speech Tagger for Arabic<sup>\*</sup>

Selçuk Köprü

Teknoloji Yazılımevi, Ltd.  
METU Technopolis  
06531, Ankara, TR  
selcuk.kopru@tyazilimevi.com

**Abstract.** In this paper, we present an efficient part-of-speech (POS) tagger for Arabic which is based on a Hidden Markov Model. We explore different enhancements to improve the baseline system. Despite the morphological complexity of Arabic our approach is a data driven approach and does not utilize any morphological analyzer or a lexicon as many other Arabic POS taggers. This makes our approach simple, very efficient and valuable to be used in real-life applications and the obtained accuracy results are still comparable to other Arabic POS taggers. In the experiments, we also thoroughly investigate different aspects of Arabic POS tagging including tag sets, prefix and suffix analyses which were not examined in detail before. Our part-of-speech tagger achieves an accuracy of 95.57% on a standard tagset for Arabic. A detailed error analysis is provided for a better evaluation of the system. We also applied the same approach on different languages like Farsi and German to show the language independent aspect of the approach. Accuracy rates on these languages are also provided.

## 1 Introduction

The continuous increase in the demand for processing Arabic faces many challenges. One important challenge at this point is the requirement to tag part-of-speech (POS) information in a given Arabic text with high accuracy and efficiently. Part-of-speech tagging is the task of classifying the words in a sentence into a set of classes. Output of POS taggers has a widespread usage in many Natural Language Processing (NLP) applications, such as Machine Translation (MT) or Information Retrieval (IR). The efficiency and accuracy of the POS tagger has usually a reasonable impact on the overall quality of the entire system. There are many studies on Statistical Machine Translation (SMT) systems that report gain in evaluation scores when a POS tagger is used [14,18,17,6].

POS tagging is the task of labeling words in an input sentence with part-of-speech and additional linguistic information. The words in the input sentence must be tokenized before the tagging process. There are two main approaches to POS tagging: rule-based tagging and stochastic tagging. Rule based taggers

---

<sup>\*</sup> This work was supported by Applications Technology, Inc. (Apptek).

# An Evaluation of Part of Speech Tagging on Written Second Language Spanish

M. Pilar Valverde Ibañez

Aichi Prefectural University  
Department of Spanish and Latin American Studies  
Kumabari, Nagakute-cho, Aichi, 480-1198, Japan  
valverde@for.aichi-pu.ac.jp

**Abstract.** With the increase in the number and size of computer learner corpora in the field of Second Language Acquisition, there is a growing need to automatically analyze the language produced by learners. However, the computational tools developed for natural language processing are generally not considered as appropriate because they are designed to treat native language. This paper analyzes the reliability of two part-of-speech taggers on second language Spanish and investigates the most frequent tagger errors and the impact of learner errors in the performance of the taggers.

## 1 Introduction

An increasing number of learner corpora are being compiled in the field of Second Language Acquisition. Such corpora, defined as electronic collections of spoken or written texts produced by foreign or second language learners [1], require linguistic annotation that enables the study of learner language in a systematic way. However, most of the work on annotating learner texts has traditionally focused on the (generally manual) annotation of errors [2], and little attention has been given to the purely linguistic annotation, irrespectively of errors. With the increase in the number and size of corpora, such annotation is becoming even more necessary to study not only misuse of words but also other aspects of learner language such as the under- and overuse of words and structures.

For these reasons, it is necessary to study how state-of-the-art natural language processing tools such as part-of-speech (PoS) taggers can be (re)used in the learner domain [3, 4]. The aim of this paper is to evaluate the performance of two PoS taggers on second language Spanish and investigate what are the most frequent tagger errors and how they are related to learner errors. For the evaluation, we have tagged a sample of 5,000 words of learner language with two Spanish PoS taggers. First, we have manually revised the PoS tags assigned by them and second, we have manually annotated the sample with learner error information. Finally, we have extracted quantitative information about the taggers and the learners' performance.

The paper is organized as follows. Section 2 deals with the type of texts and taggers used for the evaluation and makes a distinction between two types of

# Onoma: A Linguistically Motivated Conjugation System for Spanish Verbs

Luz Rello<sup>1,\*</sup> and Eduardo Basterrechea<sup>2</sup>

<sup>1</sup> NLP & Web Research Group  
Dept. of Information and Communication Technologies  
Universitat Pompeu Fabra

Barcelona, Spain

<sup>2</sup> Molino de Ideas s.a.

Nanclares de Oca, 1F

Madrid, Spain

**Abstract.** In this paper we introduce a new conjugating tool which generates and analyses both existing verbs and verb neologisms in Spanish. This application of finite state transducers is based on novel linguistically motivated morphological rules describing the verbal paradigm. Given that these transducers are simpler than the ones created in previous developments and are easy to learn and remember, the method can also be employed as a pedagogic tool in itself. A comparative evaluation of the tool against other online conjugators demonstrates its efficacy.

## 1 Introduction

Although the literature about online Spanish conjugators is scarce, it does reveal that some are fully memory based (DRAE)<sup>1</sup> while others rely on finite state morphological rules [17]<sup>2</sup>.

To the best of our knowledge, the goal of most of the work related to verbal morphology was not the creation of an end-user tool such as a conjugator. However, both machine learning and rule-based approaches have been taken into consideration when processing inflectional morphology. While instance based-learning algorithms can induce efficient morphological patterns from large training data [2,1,5,13], approaches using finite state transducers [19,8,6] do enable the implementation of robust morphological analyzer-generators which are successful in handling concatenation phenomena [4].

The Onoma conjugator<sup>3</sup> was implemented as a cascade of finite state transducers that implements a decision tree. The use of finite state transducers (FSTs)

---

\* While developing this work the first author's institution was Molino de Ideas s.a.

<sup>1</sup> Conjugator from the Dictionary of the Royal Spanish Academy (DRAE). Available at: <http://buscon.rae.es/draeI/>

<sup>2</sup> The conjugator developed by Grupo de Estructuras de Datos y Lingüística Computacional (GEDLC) at the University of Las Palmas de Gran Canaria, which is available at: [www.gedlc.ulpgc.es/investigacion/scogeme02/flexver.htm](http://www.gedlc.ulpgc.es/investigacion/scogeme02/flexver.htm)

<sup>3</sup> Developed and funded by Molino de Ideas. <http://conjugador.onoma.es>

# Measuring Similarity of Word Meaning in Context with Lexical Substitutes and Translations

Diana McCarthy

Lexical Computing Ltd.  
Brighton, BN1 6WE  
diana@dianamccarthy.co.uk

**Abstract.** Representation of word meaning has been a topic of considerable debate within the field of computational linguistics, and particularly in the subfield of word sense disambiguation. While word senses enumerated in manually produced inventories have been very useful as a start point to research, we know that the inventory should be selected for the purposes of the application. Unfortunately we have no clear understanding of how to determine the appropriateness of an inventory for monolingual applications, or when the target language is unknown in cross-lingual applications. In this paper we examine datasets which have paraphrases or translations as alternative annotations of lexical meaning on the same underlying corpus data. We demonstrate that overlap in lexical paraphrases (substitutes) between two uses of the same lemma correlates with overlap in translations. We compare the degree of overlap with annotations of usage similarity on the same data and show that the overlaps in paraphrases or translations also correlate with the similarity judgements. This bodes well for using any of these methods to evaluate unsupervised representations of lexical semantics. We do however find that the relationship breaks down for some lemmas, but this behaviour on a lemma by lemma basis itself correlates with low inter-tagger agreement and higher proportions of mid-range points on a usage similarity dataset. Lemmas which have many inter-related usages might potentially be predicted from such data.

## 1 Introduction

Words mean different things in different contexts and if we want systems to interpret and produce language as humans do then we need to build systems that can handle this. Work in computational lexical semantics has been dominated by work involving predefined inventories of senses, particularly in the subfield of word sense disambiguation (WSD). The use of inventories such as WordNet [1] have been a major catalyst for work in lexical semantics and certainly there are good reasons why one might want to use such an inventory, for example to exploit the other information contained therein. However, frequently inventories are used because that is how the gold-standard data has been annotated, rather

# A Quantitative Evaluation of Global Word Sense Induction

Marianna Apidianaki and Tim Van de Cruys

Alpage, INRIA & University Paris 7

Domaine de Voluceau

Rocquencourt, B.P. 105

F-78153 Le Chesnay cedex

{Marianna.Apidianaki,Tim.Van\_de\_Cruys}@inria.fr

**Abstract.** Word sense induction (WSI) is the task aimed at automatically identifying the senses of words in texts, without the need for hand-crafted resources or annotated data. Up till now, most WSI algorithms extract the different senses of a word ‘locally’ on a per-word basis, i.e. the different senses for each word are determined separately. In this paper, we compare the performance of such algorithms to an algorithm that uses a ‘global’ approach, i.e. the different senses of a particular word are determined by comparing them to, and demarcating them from, the senses of other words in a full-blown word space model. We adopt the evaluation framework proposed in the SemEval-2010 Word Sense Induction & Disambiguation task. All systems that participated in this task use a local scheme for determining the different senses of a word. We compare their results to the ones obtained by the global approach, and discuss the advantages and weaknesses of both approaches.

## 1 Introduction

Word sense induction (WSI) methods automatically identify the senses of words in texts, without the need for predefined resources or annotated data. These methods offer an alternative to the use of expensive hand-crafted resources developed according to the ‘fixed list of senses’ paradigm, which present several drawbacks for efficient semantic processing [1]. The assumption underlying unsupervised WSI methods is the distributional hypothesis of meaning [2], according to which words that occur in similar contexts tend to be similar. In distributional semantic analysis, the co-occurrences of words in texts constitute the features that serve to calculate their similarity. Following this approach, data-driven WSI algorithms calculate the similarity of the contexts of polysemous target words and group them into clusters. The resulting clusters describe the target word senses.

The unsupervised algorithms used for WSI can be distinguished into *local* and *global*. Local algorithms work on a per-word basis, determining the senses for each word separately. Algorithms that use a global approach determine the different senses of a particular word by comparing them to, and demarcating them from, the senses of other words in a full-blown word space model.

# Incorporating Coreference Resolution into Word Sense Disambiguation

Shangfeng Hu and Chengfei Liu

Faculty of Information and Communication Technologies  
Swinburne University of Technology,  
Hawthorn, VIC 3122, Australia  
{shu, cliu}@groupwise.swin.edu.au

**Abstract.** Word sense disambiguation (WSD) and coreference resolution are two fundamental tasks for natural language processing. Unfortunately, they are seldom studied together. In this paper, we propose to incorporate the coreference resolution technique into a word sense disambiguation system for improving disambiguation precision. Our work is based on the existing instance knowledge network (IKN) based approach for WSD. With the help of coreference resolution, we are able to connect related candidate dependency graphs at the candidate level and similarly the related instance graph patterns at the instance level in IKN together. Consequently, the contexts which can be considered for WSD are expanded and precision for WSD is improved. Based on Senseval-3 all-words task, we run extensive experiments by following the same experimental approach as the IKN based WSD. It turns out that each combined algorithm between the extended IKN WSD algorithm and one of the best five existing algorithms consistently outperforms the corresponding combined algorithm between the IKN WSD algorithm and the existing algorithm.

**Keywords:** Word sense disambiguation, coreference resolution, natural language processing.

## 1 Introduction

Word sense disambiguation (WSD) is one of the core research topics of natural language processing for identifying which sense of a word is used in a sentence, when the word has multiple meanings. It remains as an open problem in natural language processing and has important applications in areas such as machine translation, knowledge acquisition, and information retrieval [1].

Supervised WSD approaches provide the state of the art performances in benchmark evaluation [2]. Decadt et al. [3] proposed a memory-based approach which provides the best performance in senseval-3 all word tasks. Unsupervised WSD approaches were also proposed because manual supervision is a cost heavy task. Some WSD systems are built on lexical knowledge base [4-9]. Navigli [20] also proposed an integration of a knowledge-based system to improve supervised systems. They explore and calculate the semantic relationships between concepts in semantic

# Deep Semantics for Dependency Structures

Paul Bédaride<sup>1</sup> and Claire Gardent<sup>2</sup>

<sup>1</sup> Universität Stuttgart  
paul.bedaride@loria.fr  
<sup>2</sup> CNRS/LORIA  
claire.gardent@loria.fr

**Abstract.** Although dependency parsers have become increasingly popular, little work has been done on how to associate dependency structures with deep semantic representations. In this paper, we propose a semantic calculus for dependency structures which can be used to construct deep semantic representations from joint syntactic and semantic dependency structures similar to those used in the ConLL 2008 Shared Task.

**Keywords:** Dependency graphs, Deep Semantics, Graph Rewriting.

## 1 Introduction

Deep semantics have been developed for stochastic categorial parsers [1] and for parsers based on phrase structure grammars [2, 3, 4]. Much less work has been done, however, on combining dependency parsers with a deep semantics calculus. Although [5] sketches a syntax-semantics interface for dependency grammar, the proposed approach requires a constraint-based, tightly interleaved construction of dependency, predicate/argument and scoping structure which is not easily adaptable to the output of contemporary dependency parsers. Similarly, [6] presents a formalism for semantic construction from dependency structures. However, the approach incorrectly assumes that semantic dependencies match syntactic dependencies and so fails to generalise (cf. Section 2).

In this paper, we present an approach for rewriting dependency graphs into deep semantic representations that can be applied to joint syntactic and semantic dependency structures similar to those used in the ConLL 2008 Shared Task. We start by discussing a number of issues raised by dependency structures in relation to semantic construction and by motivating the choices underlying our approach (Section 2). We then present our proposal (Section 3).

## 2 Motivations

In essence, a dependency structure consists of nodes labelled with lexical items (and optionally, parts-or-speech) and linked by binary asymmetric relations called dependencies. Figure 5 illustrates this with the plain (non bold) nodes and edges forming a possible dependency graph for the sentence “John seems to love Mary”.

# Combining Heterogeneous Knowledge Resources for Improved Distributional Semantic Models

György Szarvas\*, Torsten Zesch, and Iryna Gurevych

Ubiquitous Knowledge Processing Lab, Computer Science Department,  
Technische Universität Darmstadt,  
Hochschulstr. 10, D-64289 Darmstadt, Germany  
<http://www.ukp.tu-darmstadt.de>

**Abstract.** The Explicit Semantic Analysis (ESA) model based on term cooccurrences in Wikipedia has been regarded as state-of-the-art semantic relatedness measure in the recent years. We provide an analysis of the important parameters of ESA using datasets in five different languages. Additionally, we propose the use of ESA with multiple lexical semantic resources thus exploiting multiple evidence of term cooccurrence to improve over the Wikipedia-based measure. Exploiting the improved robustness and coverage of the proposed combination, we report improved performance over single resources in word semantic relatedness, solving word choice problems, classification of semantic relations between nominals, and text similarity.

## 1 Introduction

Semantic relatedness (SR) aims at measuring how related the meaning, i.e. the semantic content of two words is. Computing the SR of words finds applications in many classical Natural Language Processing (NLP) problems like Word Sense Disambiguation [24], Information Retrieval [29,22], Cross-Language Information Retrieval [5], Text Categorization [10], Information Extraction [26], Coreference Resolution [27], or Spelling Error Detection [3].

Most of the SR measures proposed in the past have two limitations. First, they only exploit the implicit knowledge encoded in *a single* structured knowledge source like WordNet or Wikipedia, or a large text collection like the World Wide Web, but do not exploit the complementary knowledge in multiple resources through combination. Second, most measures are designed to compute relatedness between words, not between longer text segments. However, SR has important applications both on the word level (Word Sense Disambiguation, Spelling Error Correction), and on the text level (Information Retrieval, Text Categorization). Therefore, in this paper, we address the combination of the knowledge encoded in heterogeneous, independent knowledge resources to obtain better and more robust performance paying attention to direct applicability to word pairs and pairs of texts alike.

---

\* On leave from Research Group on Artificial Intelligence of the Hungarian Academy of Sciences.

# Improving Text Segmentation with Non-systematic Semantic Relation

Viet Cuong Nguyen, Le Minh Nguyen, and Akira Shimazu

School of Information Science  
Japan Advanced Institute of Science and Technology  
1-1 Asahidai, Nomi, Ishikawa 923-1211, Japan  
{cuongnv, nguyenml, shimazu}@jaist.ac.jp

**Abstract.** Text segmentation is a fundamental problem in natural language processing, which has application in information retrieval, question answering, and text summarization. Almost previous works on unsupervised text segmentation are based on the assumption of lexical cohesion, which is indicated by relations between words in the two units of text. However, they only take into account the reiteration, which is a category of lexical cohesion, such as word repetition, synonym or superordinate. In this research, we investigate the non-systematic semantic relation, which is classified as collocation in lexical cohesion. This relation holds between two words or phrases in a discourse when they pertain to a particular theme or topic. This relation has been recognized via a topic model, which is, in turn, acquired from a large collection of texts. The experimental results on the public dataset show the advantages of our approach in comparison to the available unsupervised approaches.

**Keywords:** text segmentation, lexical cohesion, topic modeling.

## 1 Introduction

Text segmentation is one of the fundamental problems in natural language processing with applications in information retrieval, text summarization, information extraction, etc. [15]. It is a process of splitting a document or a continuous stream of text into topically coherent segments. Text segmentation methods can be divided into two categories, by the structure of the output that is linear segmentation [4,7,11,14,16,22,24] and hierarchical segmentation [20], or by the algorithms that are unsupervised segmentation or supervised segmentation. In this research, we focus on the unsupervised-linear text segmentation method. The main advantage of unsupervised approach is that it does not require labeled data and is domain independent.

Almost unsupervised text segmentation methods are based on the assumption of cohesion [10], which is a device for making connections between parts of the text. Cohesion is achieved through the use of reference, substitution, ellipsis, conjunction, and lexical cohesion. The most frequent type is lexical cohesion, which is created by using semantically related words. Halliday and Hasan, in

# Automatic Identification of Cause-Effect Relations in Tamil Using CRFs

Menaka S., Pattabhi R.K. Rao, and Sobha Lalitha Devi

AU-KBC Research Centre, MIT Campus of Anna Univeristy,  
Chennai, India

{menakas , pattabhi , sobha}@au-kbc.org

**Abstract.** We present our work on automatic identification of cause-effect relations in a given Tamil text. Based on the analysis of causal constructions in Tamil, we identified a set of causal markers for Tamil and arrived at certain features used to develop our language model. We manually annotated a Tamil corpus of 8648 sentences for cause-effect relations. With this corpus, we developed the model for identifying causal relations using the machine learning technique, Conditional Random Fields (CRFs). We performed experiments and the results are encouraging. We performed an error analysis of the results and found that the errors can be attributed to some very interesting structural interdependencies between closely occurring causal relations. After comparing these structures in Tamil and English, we claim that at discourse level, the complexity of structural interdependencies between causal relations is more complex in Tamil than in English due to the free word order nature of Tamil.

**Keywords:** Cause and Effect; CRFs, Tamil; discourse; structural interdependency; machine learning.

## 1 Introduction

The analysis and modeling of discourse structure has been an important area of linguistic research in recent times and it is indeed crucial for building efficient Natural language processing (NLP) applications. The automatic identification and extraction of discourse relations can improve the performance of NLP applications like Question Answering and Information Extraction. One such discourse relation, the causal relation is the focus of this paper.

Extractions of Causal relations in English [3, 4] and in other languages like Thai [11] have been attempted by researchers from the Data mining or Knowledge Acquisition perspective. Some researchers [9] have focused on recognition of discourse relations using cue phrases or detection of implicit discourse relations, but not extraction of arguments. Others [2, 18] have tried identification of the arguments of all discourse connectives in the PDTB.

Our work aims at extraction of causal relations from a text comprehension perspective i.e., we're interested in what is expressed in text rather than what is a causal relation in the real world. Our objective is to identify causal markers and the text spans of their two arguments. Our work is closer to [2] and [18] in identification

# Comparing Approaches to Tag Discourse Relations

Shamima Mithun and Leila Kosseim

Concordia University,  
Department of Computer Science and Software Engineering,  
Montreal, Quebec, Canada  
{s\_mithun,kosseim}@encs.concordia.ca

**Abstract.** It is widely accepted that in a text, sentences and clauses cannot be understood in isolation but in relation with each other through discourse relations that may or may not be explicitly marked. Discourse relations have been found useful in many applications such as machine translation, text summarization, and question answering; however, they are often not considered in computational language applications because domain and genre independent robust discourse parsers are very few. In this paper, we analyze existing approaches to identify five discourse relations automatically (namely, *comparison*, *contingency*, *illustration*, *attribution*, and *topic-opinion*), and propose a new approach to identify *attributive* relations. We evaluate the accuracy of each approach with respect to the discourse relations it can identify and compare it to a human gold standard. The evaluation results show that the state of the art systems are rather effective at identifying most of the relations considered, but other relations such as *attribution* are still not identified with high accuracy.

## 1 Introduction

It is widely accepted that sentences and clauses in a text cannot be understood in isolation but in relation with each other. A text is not a linear combination of clauses but a hierarchical organized group of clauses placed together based on informational and interactional relations to one another. For example, in the sentence “*If you want the full Vista experience, you’ll want a heavy system and graphics hardware, and lots of memory*”, the first and second clauses do not bear much meaning independently; they become more meaningful when we realize that they are related through the discourse relation *condition*.

In a discourse, different kinds of relations such as *contrast*, *causality*, *elaboration* may be expressed. The use of such discourse structures modelled by rhetorical predicates (described in section 2) have been found useful in many applications such as document summarization and question answering ([9, 7]). For example, [9] showed that rhetorical predicates can be used to select the content and generate coherent text in question answering with the help of schemata. Recently, [10] has demonstrated that rhetorical predicates can be useful in blog

# Semi-supervised Discourse Relation Classification with Structural Learning

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka

Graduate School of Information Science & Technology  
The University of Tokyo

7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

hugo@mi.ci.i.u-tokyo.ac.jp, danushka@iba.t.u-tokyo.ac.jp,  
ishizuka@i.u-tokyo.ac.jp

**Abstract.** The corpora available for training discourse relation classifiers are annotated using a general set of discourse relations. However, for certain applications, custom discourse relations are required. Creating a new annotated corpus with a new relation taxonomy is a time-consuming and costly process. We address this problem by proposing a semi-supervised approach to discourse relation classification based on Structural Learning. First, we solve a set of auxiliary classification problems using unlabeled data. Second, the learned classifiers are used to extend feature vectors to train a discourse relation classifier. By defining a relevant set of auxiliary classification problems, we show that the proposed method brings improvement of at least 50% in accuracy and F-score on the RST Discourse Treebank and Penn Discourse Treebank, when small training sets of ca. 1000 training instances are employed. This is an attractive perspective for training discourse relation classifiers on domains where little amount of labeled training data is available.

## 1 Introduction

Detecting the discourse relations underlying the different units of a text is crucial for several NLP applications, such as text summarization [1] or dialogue generation [2]. To date, only three major annotated corpora are available, the RST Discourse Treebank (RSTDT) [3], the Discourse Graphbank [4], and the Penn Discourse Treebank (PDTB) [5]. The RSTDT is based on the Rhetorical Structure Theory framework (RST) [6], and annotation is done using a set of 78 fine-grained discourse relations, usually grouped by researchers into a set of 18 more general relations [3]. In the Discourse GraphBank, annotation is done using a set of 11 discourse relations. Finally, in the PDTB, annotation is done in a hierarchical fashion, with 4 relations at the highest-level, and 20 at the most detailed level.

However, in some applications, we must extract discourse relations that are different from the ones defined in above-mentioned discourse theories. In [7] for instance, it is shown that the use of a RST discourse parser improves the detection of relevant information in clinical guidelines. Notably, certain RST discourse relations such as TEMPORAL or CONSEQUENCE are useful in the context

# Integrating Japanese Particles Function and Information Structure

Akira Ohtani

Osaka Gakuin University  
Faculty of Informatics  
2-36-1 Kishibe-minami, Suita, Osaka 564-8511 Japan  
ohtani@ogu.ac.jp

**Abstract.** This paper presents a new analysis of the discourse functions of Japanese particles *wa* and *ga*. Such functions are integrated with information structures into the constraint-based grammar under the HPSG framework. We examine the distribution of these particles and demonstrate how the thematic-rhematic dichotomy of the constituent can be determined by the informational status of one or more of its daughter constituents through various linguistic constraints. We show that the relation between syntactic constituency and information structure of a sentence is not a one-to-one mapping as a purely syntax-based analysis assumes, and then propose the multi-dimensional grammar which expresses mutual constraints on the thematic-rhematic interpretation, syntax and phonology.

## 1 Introduction

Information Structure (IS) plays a crucial role for ensuring coherence in discourse. In many languages, intonation is the primary means of conveying IS. The mini-dialogue in (1), where **bold face** corresponds to the so called B-accent (L+H\*) and SMALL CAPITALS indicate the word bearing the so called A-accent (H\*), illustrates the connection between IS and accent in English.

- (1) Speaker Q: So tell me about the people in the White House.  
Anything I should know?

Speaker A<sub>1</sub>: Yes. [ $\theta$  The **president**] [ $\rho$  hates the Delft CHINA SET].  
Don't use it. (Engdahl and Vallduví [1]:5, ex.3, Modified.)

The information conveyed by a sentence is split into new information *rheme* ( $\rho$  *focus*) and information already present in the discourse *theme* ( $\theta$ , *topic*).

It has been observed that languages adopt different means to encode their IS: English employs prosody, Catalan relies on word order, and Greek uses both. In addition to those means, Japanese utilizes morphology.

- (2) Speaker A<sub>2</sub>: [ $\theta$  Daitooryoo-*wa*] [ $\rho$  maisen-no SYOKKI-*ga* okonomi desu].  
president- $\theta$  Meissen-GEN china.set-NOM like  
'The president likes the Meissen china set.'

In (2) theme and rheme are identified by the particles, *wa* and *ga*, respectively.

# Assessing Lexical Alignment in Spontaneous Direction Dialogue Data by Means of a Lexicon Network Model

Alexander Mehler<sup>1</sup>, Andy Lücking<sup>2</sup>, and Peter Menke<sup>2</sup>

<sup>1</sup> Goethe University Frankfurt am Main

mehler@em.uni-frankfurt.de

<sup>2</sup> Bielefeld University

{andy.luecking,peter.menke}@uni-bielefeld.de

**Abstract.** We apply a network model of lexical alignment, called *Two-Level Time-Aligned Network Series*, to natural route direction dialogue data. The model accounts for the structural similarity of interlocutors' dialogue lexica. As classification criterion the directions are divided into effective and ineffective ones. We found that effective direction dialogues can be separated from ineffective ones with a hit ratio of 96% with regard to the structure of the corresponding dialogue lexica. This value is achieved when taking into account just nouns. This hit ratio decreases slightly as soon as other parts of speech are also considered. Thus, this paper provides a machine learning framework for telling apart effective dialogues from insufficient ones. It also implements first steps in more fine-grained alignment studies: we found a difference in the efficiency contribution between (the interaction of) lemmata of different parts of speech.

## 1 Motivation

According to the *Interactive Alignment Model* [1, IAM], mental representations of dialogue partners on all linguistic levels become more and more similar, i.e. *aligned*, during their communicative interaction. Since the linguistic levels – phonetic, lexical, syntactic, semantic, situation model – are interconnected, alignment propagates through these levels. Via this spreading of alignment, global alignment, that is, alignment of situation models, can be a result of local alignment on lower levels. Thus, the IAM provides an account to the ease and efficiency of dialogical communication beyond explicit negotiation. Part of the efficiency of communication is the fulfillment of the dialogue task or purpose. Consequently, we would expect that *more aligned dialogues are more successful* – a proposition we make productive below.

The central mechanism that is acknowledged within the IAM is *priming*.<sup>1</sup> Priming is typically understood and modeled as spreading activation in neural networks. Two varieties of activation have to be distinguished:

---

<sup>1</sup> But see [2] for an argument that priming cannot be the process that implements alignment.

# Towards Well-Grounded Phrase-Level Polarity Analysis

Robert Remus and Christian Hänig

University of Leipzig  
Natural Language Processing Group  
Department of Computer Science  
04103 Leipzig, Germany  
{rremus, chaenig}@informatik.uni-leipzig.de

**Abstract.** We propose a new rule-based system for phrase-level polarity analysis and show how it benefits from empirically validating its polarity composition through surveys with human subjects. The system’s two-layer architecture and its underlying structure, i.e. its composition model, are presented. Two functions for polarity aggregation are introduced that operate on newly defined semantic categories. These categories detach a word’s syntactic from its semantic behavior. An experimental setup is described that we use to carry out a thorough evaluation. It incorporates a newly created German-language data set that is made freely and publicly available. This data set contains polarity annotations at word-level, phrase-level and sentence-level and facilitates comparability between different studies and reproducibility of our results.

## 1 Introduction

With the advancing integration of the Internet into our everyday life, the amount of user generated content grows rapidly. People blog about their experiences, discuss in fora, author product reviews or twitter short messages. They do not stick to certain topics but write about everything of interest, e.g. holidays or recent purchases. In “Web 2.0”, people express their *opinion* directly and frankly without being asked to do so and hence, this content has an immeasurable value for market research. While for marketing purposes, sentence- or even document-level analysis may suffice, a more *fine-grained analysis* is essential for deeper investigations established in business environments (e.g. quality assurance, competition analysis).

Whereas most approaches to sentiment analysis focus on two- or three-way classification of words (cf. [1]), sentences (cf. [2,3]) or complete documents (cf. [4]), with both rule-based or machine learning techniques, they all make a general assumption: each sentence or document deals with exactly one *topic*. This should be true for most product reviews, but within fora discussions or blog entries, people often address multiple topics. Typical sentences like *Ich mag X, aber Y sieht komisch aus.* (*I like X, but Y looks strange.*) contain more than one topic. Thus, in order to perform a thorough and fine-grained analysis, one has to delve

# Implicit Feature Identification via Co-occurrence Association Rule Mining

Zhen Hai, Kuiyu Chang, and Jung-jae Kim

School of Computer Engineering, Nanyang Technological University, Singapore  
{haiz0001, askychang, jungjae.kim}@ntu.edu.sg

**Abstract.** In sentiment analysis, identifying features associated with an opinion can help produce a finer-grained understanding of online reviews. The vast majority of existing approaches focus on explicit feature identification, few attempts have been made to identify implicit features in reviews. In this paper, we propose a novel two-phase co-occurrence association rule mining approach to identifying implicit features. Specifically, in the first phase of rule generation, for each opinion word occurring in an explicit sentence in the corpus, we mine a significant set of association rules of the form [opinion-word, explicit-feature] from a co-occurrence matrix. In the second phase of rule application, we first cluster the rule consequents (explicit features) to generate more robust rules for each opinion word mentioned above. Given a new opinion word with no explicit feature, we then search a matched list of robust rules, among which the rule having the feature cluster with the highest frequency weight is fired, and accordingly, we assign the representative word of the cluster as the final identified implicit feature. Experimental results show considerable improvements of our approach over other related methods including baseline dictionary lookups, statistical semantic association models, and bi-bipartite reinforcement clustering.

**Keywords:** opinion mining, opinion word, implicit feature, co-occurrence, association rule.

## 1 Introduction

With the rapid development of the Internet, a large amount of subjective user-generated content is available in online forums, blogs, and shopping websites. The task of subjectivity classification is to recognize opinionated sentences or documents apart from the text segments that show objective information [1–5]. Meanwhile, the task of sentiment identification primarily deals with classifying sentiment orientations expressed in text [6–9]. Document-level sentiment analysis, or opinion mining, can classify the overall subjectivity or sentiment orientation expressed in the review content, but fails to infer the sentiment associated with individual features. This problem also happens, though to a lesser extent, in the sentence-level sentiment analysis, as shown in the following cell phone review example:

# Construction of Wakamono Kotoba Emotion Dictionary and Its Application

Kazuyuki Matsumoto and Fuji Ren

Faculty of Engineering, University of Tokushima  
2-1 Minami-josanjima, Tokushima, 770-8506, Japan  
{matumoto,ren}@is.tokushima-u.ac.jp

**Abstract.** Currently, we can find a lot of weblogs written by young people. In the weblogs, they tend to describe their undiluted emotions or opinions. To analyze the emotions of young people and what causes those emotions, our study focuses on the specific Japanese language used among young people, which is called *Wakamono Kotoba*. The proposed method classifies *Wakamono Kotoba* into emotion categories based on superficial information and the descriptive texts of the words. Specifically, the method uses literal information used for *Wakamono Kotoba*, such as Katakana, Hiragana, and Kanji, etc., stroke count, and the difficulty level of the Kanji as features. Then we classified *Wakamono Kotoba* into emotion categories according to the superficial similarity between the word and the *Wakamono Kotoba* registered in the dictionary with an annotation of its emotional strength level. We also proposed another method to classify *Wakamono Kotoba* into emotion categories by using the co-occurrence relation between the words included in the descriptive text of the *Wakamono Kotoba* and the emotion words included in the existing emotion word dictionary. We constructed the *Wakamono Kotoba* emotion dictionary based on these two methods. Finally, the applications of the *Wakamono Kotoba* emotion dictionary are discussed.

**Keywords:** Wakamono kotoba, emotion dictionary, emotion corpus.

## 1 Introduction

Currently, a variety of languages are used to convey text information on the Internet. Internet terminology is an example and it is usually used only on the Internet. Another example is *Wakamono Kotoba*, a Japanese language used by the younger generation ranging from teenagers to those in their late 20's. They are also frequently used on the Internet. Without processing *Wakamono Kotoba* correctly, it would be difficult to extract the living opinions of younger people from the information on the Web. To precisely recognize *Wakamono Kotoba* in a sentence, it is necessary to register these words in the dictionary. However, the process is not efficient.

Therefore, technology to automatically recognize unknown *Wakamono Kotoba* words from a sentence and to analyze such words semantically would be useful.

# Temporal Analysis of Sentiment Events – A Visual Realization and Tracking

Dipankar Das<sup>1</sup>, Anup Kumar Kolya<sup>1</sup>, Asif Ekbal<sup>2</sup>, and Sivaji Bandyopadhyay<sup>1</sup>

<sup>1</sup> Department of Computer Science and Engineering, Jadavpur University,  
Kolkata 700 032, India

<sup>2</sup> Department of Information Engineering and Computer Science,  
University of Trento, Italy  
{dipankar.dipnil2005, anup.kolya, asif.ekbal}@gmail.com,  
sivaji\_cse\_ju@yahoo.com

**Abstract.** In recent years, extraction of temporal relations for events that express sentiments has drawn great attention of the Natural Language Processing (NLP) research communities. In this work, we propose a method that involves the association and contribution of sentiments in determining the event-event relations from texts. Firstly, we employ a machine learning approach based on Conditional Random Field (CRF) for solving the problem of Task C (identification of event-event relations) of TempEval-2007 within TimeML framework by considering *sentiment* as a feature of an event. Incorporating sentiment property, our system achieves the performance that is better than all the participated state-of-the-art systems of TempEval 2007. Evaluation results on the Task C test set yield the F-score values of 57.2% under the strict evaluation scheme and 58.6% under the relaxed evaluation scheme. The positive or negative coarse grained sentiments as well as the Ekman's six basic universal emotions (or, fine grained sentiments) are assigned to the events. Thereafter, we analyze the temporal relations between events in order to track the sentiment events. Representation of the temporal relations in a graph format shows the shallow visual realization path for tracking the sentiments over events. Manual evaluation of temporal relations of sentiment events identified in 20 documents sounds satisfactory from the purview of event-sentiment tracking.

**Keywords:** Temporal Relations, CRF, TempEval-2007, TimeML, Sentiment Event, Visual Tracking.

## 1 Introduction

The kinds of states which change and thus might need to be located in time are referred as events in the present context. The event entities are represented by finite clauses, nonfinite clauses, nominalizations, event-referring nouns, adjectives and even some kinds of adverbial clauses. In general, the events are described in different newspaper texts, stories and other important documents where occurrence time of events, temporal location and ordering of the events are specified. Several earlier methods have been proposed for detecting and tracking events from text archives [1].

# Highly-Inflected Language Generation Using Factored Language Models

Eder Miranda de Novais, Ivandré Paraboni, and Diogo Takaki Ferreira

School of Arts, Sciences and Humanities, University of São Paulo (USP / EACH)  
Av. Arlindo Bettio, 1000 - São Paulo, Brazil  
{eder.novais, ivandre, diogo.ferreira}@usp.br

**Abstract.** Statistical language models based on n-gram counts have been shown to successfully replace grammar rules in standard 2-stage (or ‘generate-and-select’) Natural Language Generation (NLG). In highly-inflected languages, however, the amount of training data required to cope with n-gram sparseness may be simply unobtainable, and the benefits of a statistical approach become less obvious. In this work we address the issue of text generation in a highly-inflected language by making use of factored language models (FLM) that take morphological information into account. We present a number of experiments involving the use of simple FLMS applied to various surface realisation tasks, showing that FLMS may implement 2-stage generation with results that are far superior to standard n-gram models alone.

**Keywords:** Text Generation, Surface Realisation, Language Modelling.

## 1 Introduction

In Natural Language Generation (NLG) systems, surface realisation can be viewed as the task that takes as an input a set of features representing a sentence specification (or *what* to say), and produces the corresponding output string (*how* to say it) [1]. In the case of symbolic generation, the input to surface realisation will normally have to be provided in a high level of detail, making the surface realisation module<sup>1</sup> difficult to adapt to applications that are not linguistically-motivated by design.

By contrast, with the more recent use of statistical methods in NLG, the issue of input specification to surface realisation has become in many ways more manageable. Standard 2-stage (or ‘generate-and-select’) architectures as in [3]<sup>2</sup> complement an underspecified input by overgenerating a large number of alternative realisations (often including ungrammatical sentences) and selecting the most likely output according to a statistical language model.

Language models may however suffer from data sparseness, and statistical NLG relies heavily on large corpora as training data. Early work in the field [3]

---

<sup>1</sup> See [2] for details on a typical NLG architecture.

<sup>2</sup> Given the limited availability of NLP resources for our target language, in this paper we do not discuss state-of-art grammar acquisition approaches such as [4].

# Prenominal Modifier Ordering in Bengali Text Generation

Sumit Das, Anupam Basu\*, and Sudeshna Sarkar

Indian Institute of Technology, Kharagpur  
Department of Computer Science and Engineering  
Kharagpur, West Bengal, India – 721302  
{sumitdas, anupam, sudeshna}@cse.iitkgp.ernet.in

**Abstract.** In this paper, we propose a machine learning based approach for ordering adjectival premodifiers of a noun phrase (NP) in Bengali. We propose a novel method to learn the pairwise orders of the modifiers. Using the learned pairwise orders, longer sequences of pronominal modifiers are ordered following a graph based method. The proposed modifier ordering approach is compared with an existing approach using our own dataset. We have achieved approximately 4% increment in the *F-measure* with our approach indicating an overall improvement. The modifier ordering approach proposed here can be implemented in a Bengali text generation system resulting in more fluent and natural output.

## 1 Introduction

Natural Language Generation (NLG) systems should produce text which is meaning-preserving and fluent. Ordering pronominal modifiers is an important task in text generation, because wrongly ordered modifiers in NPs affect the meaning and fluency of the generated text. In English, pronominal modifiers can occur almost in any order, depending on the context. Some orders are more marked than the others, but strictly speaking none are ungrammatical. For example, for the NP in (1), (1a) is more fluent than the other two orders.

1. (a) *charming young blond lady*
- (b) \* *blond young charming lady*<sup>1</sup>
- (c) \* *blond charming young lady*

Though there exists some consensus that the pronominal modifier ordering is partly governed by the semantic constraints, but the exact semantic constraints are not known. Few early studies [1,2] manually analyzed small corpora, based on which they placed the modifiers in broad semantic classes. They defined rules to impose order among the modifier classes. The modifiers were ordered according to the order of the classes to which they belong. The recent works on

---

\* Prof. Anupam Basu is the Director (Hon.) of Society for Natural Language Technology Research (SNLTR), Kolkata, West Bengal, India.

<sup>1</sup> The ‘\*’ marked NPs are not fluent.

# Bootstrapping Multiple-Choice Tests with THE-MENTOR

Ana Cristina Mendes, Sérgio Curto, and Luísa Coheur

Spoken Language Systems Laboratory - L<sup>2</sup>F/INESC-ID  
Instituto Superior Técnico, Technical University of Lisbon  
R. Alves Redol, 9 - 2<sup>o</sup> - 1000-029 Lisboa, Portugal  
{ana.mendes, sergio.curto, luisa.coheur}@l2f.inesc-id.pt

**Abstract.** It is very likely that, at least once in their lifetime, everyone has answered a multiple-choice test. Multiple-choice tests are considered an effective technique for knowledge assessment, requiring a short response time and with the possibility of covering a broad set of topics. Nevertheless, when it comes to their creation, it can be a time-consuming and labour-intensive task. Here, the generation of multiple-choice tests aided by computer can reduce these drawbacks: to the human assessor is attributed the final task of approving or rejecting the generated test items, depending on their quality.

In this paper we present THE-MENTOR, a system that employs a fully automatic approach to generate multiple-choice tests. In a first offline step, a set of lexico-syntactic patterns are bootstrapped by using several question/answer seed pairs and leveraging the redundancy of the Web. Afterwards, in an online step, the patterns are used to select sentences in a text document from which answers can be extracted and the respective questions built. In the end, several filters are applied to discard low quality items and distractors are named entities that comply with the question category, extracted from the same text.

## 1 Introduction

Multiple-choice tests are an effective technique for knowledge assessment, requiring a short response time and with the possibility of covering a broad set of topics. Typically, these tests consist in a number of test items, each composed by two parts: a question and a group of suggested answers. Respondents are supposed to identify the correct answer among the incorrect ones (called distractors). The following is an example of a multiple-choice test item with one correct answer and two distractors:

- Q.** *“What is the largest ocean?”*
- 
1. *Atlantic* (distractor)
  2. *Pacific* (correct answer)
  3. *Indian* (distractor)

The manual creation of multiple-choice test items is a time consuming trial and error process; in this context, computer aided multiple-choice tests generation can help reducing the amount of time allocated to this task.

# Ontology Based Interlingua Translation

Leonardo Lesmo, Alessandro Mazzei, and Daniele P. Radicioni

University of Turin, Computer Science Department  
Corso Svizzera 185, 10149 Turin  
{lesmo,mazzei,radicion}@di.unito.it

**Abstract.** In this paper we describe an interlingua translation system from Italian to Italian Sign Language. The main components of this systems are a broad coverage dependency parser, an ontology based semantic interpreter and a grammar-based generator: we provide the description of the main features of these components.

## 1 Introduction

In this paper we describe some features of a system designed to translate from Italian to Italian Sign Language (henceforth LIS). Many approaches have been proposed for automatic translation, which require different kinds of linguistic analysis. For instance, the *direct* translation paradigm requires just morphological analysis of the source sentence, while the *transfer* translation paradigm requires syntactic (and sometimes semantic) analysis too [1]. In contrast, our architecture adheres to the *interlingua* translation paradigm, i.e. it performs a deep linguistic processing in each phase of the translation, i.e. (1) deep syntactic analysis of the Italian source sentence, (2) semantic interpretation, and (3) generation in LIS of the target LIS sentence. These three phases form a pipeline of processing: the syntactic tree produced in the first phase is the input for the second phase, i.e semantic interpretation; similarly, the semantic structure produced in the second phase is the input of the third phase, i.e. generation. In order to work properly, Interlingua pipeline requires good performances in each phase of the translation. Moreover, since the semantic interpretation is crucially related to the world knowledge, the state-of-the-art computational linguistic techniques allow the interlingua approach to work only on limited domain [1]. In our work, we concentrate on the classical domain of weather forecasts.

A challenging requirement of our project is related to the target language, the LIS, that does not have a *natural* written form (which is typical of the signed languages). In our project we developed an *artificial* written form for LIS: this written form encodes the main morphological features of the signs as well as a number of non-manual features, as the gaze or the tilt of the head. Anyway, for sake of clarity in this paper we report a LIS sentence just as a sequence of *GLOSSAS*, that is the sequence of the names<sup>1</sup> of the signs, without any extra-lexical feature.

---

<sup>1</sup> A name for a sign is just a *code* necessary to represent the sign. As it is customary in the sign languages literature, we use names for the signs that are related to their rough translation into another language, Italian in our work.

# Phrasal Equivalence Classes for Generalized Corpus-Based Machine Translation

Rashmi Gangadharaiah, Ralf D. Brown, and Jaime Carbonell

Carnegie Mellon University  
{rgangadh, ralf, jgc}@cs.cmu.edu

**Abstract.** Generalizations of sentence-pairs in Example-based Machine Translation (EBMT) have been shown to increase coverage and translation quality in the past. These template-based approaches (G-EBMT) find common patterns in the bilingual corpus to generate generalized templates. In the past, patterns in the corpus were found by only few of the following ways: finding similar or dissimilar portions of text in groups of sentence-pairs, finding semantically similar words, or use dictionaries and parsers to find syntactic correspondences. This paper combines all the three aspects for generating templates. In this paper, the boundaries for aligning and extracting members (phrase-pairs) for clustering are found using chunkers (hence, syntactic information) trained independently on the two languages under consideration. Then semantically related phrase-pairs are grouped based on the contexts in which they appear. Templates are then constructed by replacing these clustered phrase-pairs by their class labels. We also perform a filtration step by simulating human labelers to obtain only those phrase-pairs that have high correspondences between the source and the target phrases that make up the phrase-pairs. Templates with English-Chinese and English-French language pairs gave significant improvements over a baseline with no templates.

**Keywords:** Generalized Example-based Machine Translation (G-EBMT), Template Induction, Unsupervised Clustering, data sparsity.

## 1 Introduction

Templates are generalizations of sentence-pairs formed by replacing sequences of words by variables. Like other data-driven MT approaches such as Statistical MT (SMT), EBMT also requires large amounts of data to perform well. Generalization was introduced in EBMT to increase coverage and improve quality in data-sparse conditions [12,4]. If the following sentence-pair (SP: source and its corresponding target sentence) is present in the bilingual training corpus and equivalence classes  $C_1$  and  $C_2$  are among the clusters available (either obtained automatically or from a bilingual speaker),

(SP) source sentence: flood prevention and development plans must  
also be drawn up for the major river basins  
target sentence: 各大流域也要制定防洪、开发治理规划

$C_1$ :

flood prevention and development plans ↔ 防洪、开发治理规划  
action plans ↔ 动预案  
emergency plans ↔ 应急预案

# A Multi-view Approach for Term Translation Spotting

Raphaël Rubino and Georges Linares

Laboratoire Informatique d'Avignon  
339, chemin des Meinajaries, BP 91228  
84911 Avignon Cedex 9, France

{raphael.rubino,georges.linares}@univ-avignon.fr

**Abstract.** This paper presents a multi-view approach for term translation spotting, based on a bilingual lexicon and comparable corpora. We propose to study different levels of representation for a term: the context, the theme and the orthography. These three approaches are studied individually and combined in order to rank translation candidates. We focus our task on French-English medical terms. Experiments show a significant improvement of the classical context-based approach, with a F-score of 40.3% for the first ranked translation candidates.

**Keywords:** Multilingualism, Comparable Corpora, Topic Model.

## 1 Introduction

Bilingual term spotting is a popular task which can be used for bilingual lexicon construction. This kind of resource is particularly useful in many Natural Language Processing (NLP) tasks, for example in cross-lingual information retrieval or Statistical Machine Translation (SMT). Some works in the literature are based on the use of bilingual parallel texts, which are often used in SMT for building translation tables [1,2]. However, the lack of parallel texts is still an issue, and the NLP community tends to use a forthcoming bilingual resource in order to build bilingual lexicons: bilingual comparable corpora.

One of the main approaches using non-parallel corpora is based on the assumption that a term and its translation share context similarities. It can be seen as a co-occurrence or a context-vector model, which depends on the lexical environment of terms [3,4]. This approach stands on the use of a bilingual lexicon, also known as bilingual seed-words. These words are used as anchor points in the source and the target language. This representation of the environment of a term has to be invariant from a language to another in order to spot correct translations. The efficiency of this approach depends on context-vectors accuracy. Authors have studied different measures between terms, variations on the context size, and similarity metrics between context-vectors [5,6,7].

In addition to context information, heuristics are often used to improve the general accuracy of the context-vector approach, like orthographic similarities

# ICE-TEA: In-Context Expansion and Translation of English Abbreviations

Waleed Ammar, Kareem Darwish, Ali El Kahki, and Khaled Hafez<sup>1</sup>

Cairo Microsoft Innovation Center, Microsoft, 306 Chorniche El-Nile, Maadi, Cairo, Egypt  
{i-waamma, kareemd, t-aleka}@microsoft.com,  
hafez.khaled@gmail.com

**Abstract.** The wide use of abbreviations in modern texts poses interesting challenges and opportunities in the field of NLP. In addition to their dynamic nature, abbreviations are highly polysemous with respect to regular words. Technologies that exhibit some level of language understanding may be adversely impacted by the presence of abbreviations. This paper addresses two related problems: (1) expansion of abbreviations given a context, and (2) translation of sentences with abbreviations. First, an efficient retrieval-based method for English abbreviation expansion is presented. Then, a hybrid system is used to pick among simple abbreviation-translation methods. The hybrid system achieves an improvement of 1.48 BLEU points over the baseline MT system, using sentences that contain abbreviations as a test set.

**Keywords:** statistical machine translation, word sense disambiguation, abbreviations.

## 1 Introduction

Abbreviations are widely used in modern texts of several languages, especially English. In a recent dump of English Wikipedia,<sup>2</sup> articles contain an average of 9.7 abbreviations per article, and more than 63% of the articles contain at least one abbreviation. At sentence level, over 27% of sentences, from news articles, were found to contain abbreviations. The ubiquitous use of abbreviations is worth some attention. Abbreviations can be acronyms, such as NASA, which are pronounced as words, or initialisms, such as BBC, which are pronounced as a sequence of letters.

Often abbreviations have multiple common expansions, only one of which is valid for a particular context. For example, Wikipedia lists 17 and 15 valid expansions for IRA and IRS respectively. However, in the sentence: “*The bank reported to the IRS all withheld taxes for IRA accounts.*” IRA conclusively refers to “*individual retirement account*” and IRS refers to “*internal revenue service*”. Zahariev (2004) states that 47.97% of abbreviations have multiple expansions (at WWWAAS<sup>3</sup>)

---

<sup>1</sup> Author was an intern at Microsoft and is currently working at the IBM Technology Development Center in Cairo.

<sup>2</sup> <http://dumps.wikimedia.org/enwiki/20100312/>

<sup>3</sup> World-Wide Web Acronym and Abbreviation Server  
<http://acronyms.silmaril.ie/>

# Word Segmentation for Dialect Translation

Michael Paul, Andrew Finch, and Eiichiro Sumita

National Institute of Information and Communications Technology  
MASTAR Project  
Kyoto, Japan  
michael.paul@nict.go.jp

**Abstract.** This paper proposes an unsupervised word segmentation algorithm that identifies word boundaries in continuous source language text in order to improve the translation quality of statistical machine translation (SMT) approaches for the translation of local dialects by exploiting linguistic information of the standard language. The method iteratively learns multiple segmentation schemes that are consistent with (1) the standard dialect segmentations and (2) the phrasal segmentations of an SMT system trained on the resegmented bitext of the local dialect. In a second step multiple segmentation schemes are integrated into a single SMT system by characterizing the source language side and merging identical translation pairs of differently segmented SMT models. Experimental results translating three Japanese local dialects (Kumamoto, Kyoto, Osaka) into three Indo-European languages (English, German, Russian) revealed that the proposed system outperforms SMT engines trained on character-based as well as standard dialect segmentation schemes for the majority of the investigated translation tasks and automatic evaluation metrics.

## 1 Introduction

Spoken languages distinguish regional speech patterns, the so-called *dialects*: a variety of a language that is characteristic of a particular group of the language's speakers. A *standard dialect* (or *standard language*) is a dialect that is recognized as the "correct" spoken and written form of the language. Dialects typically differ in terms of morphology, vocabulary and pronunciation. Various methods have been proposed to measure the relatedness between dialects using phonetic distance measures [1], string distance algorithms [2,3], or statistical models [4]. Concerning data-driven natural language processing (NLP) applications, research on dialect processing focuses on the analysis and generation of dialect morphology [5], parsing of dialect transcriptions [6], spoken dialect identification [7], and machine translation [8,9,10].

For most of the above applications, explicit knowledge about the relation between the standard dialect and the local dialect is used to create local dialect language resources. In terms of morphology, certain lemmata of word forms are shared between different dialects where the usage and order of inflectional affixes might change. The creation of rules that map between dialectic variations can

# TEP: Tehran English-Persian Parallel Corpus

Mohammad Taher Pilevar<sup>1</sup>, Hesham Faili<sup>1</sup>, and Abdol Hamid Pilevar<sup>2</sup>

<sup>1</sup> Natural Language Processing Laboratory,  
University of Tehran, Iran

{t.pilevar,h.faili}@ut.ac.ir

<sup>2</sup> Faculty of Computer Engineering,  
Bu Ali Sina University, Hamedan, Iran  
pilevar@basu.ac.ir

**Abstract.** Parallel corpora are one of the key resources in natural language processing. In spite of their importance in many multi-lingual applications, no large-scale English-Persian corpus has been made available so far, given the difficulties in its creation and the intensive labors required. In this paper, the construction process of Tehran English-Persian parallel corpus (TEP) using movie subtitles, together with some of the difficulties we experienced during data extraction and sentence alignment are addressed. To the best of our knowledge, TEP has been the first freely released large-scale (in order of million words) English-Persian parallel corpus.

## 1 Parallel Corpora

Text corpus is a structured electronic source of data to be analyzed for natural language processing applications. A corpus may contain texts in a single language (monolingual corpus) or in multiple languages (multilingual corpus). Corpora are the main resources in corpus linguistics to study the language as expressed in samples or real world text. Parallel corpora are specially formatted multilingual corpora whose contents are aligned side-by-side in order to be used for comparison purpose.

While there are various resources such as newswires, books and websites that can be used to construct monolingual corpora, parallel corpora need more specific types of multilingual resources which are comparatively more difficult to obtain. As a result, large-scale parallel corpora are rarely available especially for lesser studied languages like Persian.

### 1.1 Properties of Parallel Corpora

Parallel corpora possess some properties that should be taken into account in their development [1]. The first feature is the structural distance between the text pair which indicates whether the translation is literal or free. Literal and free translations are two basic skills of human translation. A literal translation (also known as word-for-word translation) is a translation that closely follows the form of source language. It is admitted in the machine translation community that the training data of literal type better suits statistical machine translation (SMT) systems at their present level of intelligence [2].

# Effective Use of Dependency Structure for Bilingual Lexicon Creation

Daniel Andrade<sup>1</sup>, Takuya Matsuzaki<sup>1</sup>, and Jun'ichi Tsujii<sup>2,3</sup>

<sup>1</sup> Department of Computer Science, University of Tokyo, Tokyo, Japan

<sup>2</sup> School of Computer Science, University of Manchester, Manchester, UK

<sup>3</sup> National Centre for Text Mining, Manchester, UK

{daniel.andrade,matuzaki,tsujii}@is.s.u-tokyo.ac.jp

**Abstract.** Existing dictionaries may be effectively enlarged by finding the translations of single words, using comparable corpora. The idea is based on the assumption that similar words have similar contexts across multiple languages. However, previous research suggests the use of a simple bag-of-words model to capture the lexical context, or assumes that sufficient context information can be captured by the successor and predecessor of the dependency tree. While the latter may be sufficient for a close language-pair, we observed that the method is insufficient if the languages differ significantly, as is the case for Japanese and English. Given a query word, our proposed method uses a statistical model to extract relevant words, which tend to co-occur in the same sentence; additionally our proposed method uses three statistical models to extract relevant predecessors, successors and siblings in the dependency tree. We then combine the information gained from the four statistical models, and compare this lexical-dependency information across English and Japanese to identify likely translation candidates. Experiments based on openly accessible comparable corpora verify that our proposed method can increase Top 1 accuracy statistically significantly by around 13 percentage points to 53%, and Top 20 accuracy to 91%.

## 1 Introduction

Even for resource rich languages like Japanese and English, where there are comprehensive dictionaries already available, it is necessary to constantly update those existing dictionaries to include translations of new words. As such, it is helpful to assist human translators by automatically extracting plausible translation candidates from comparable corpora. The term comparable corpora refers to a pair of non-parallel corpora written in different languages, but which are roughly about a similar topic. The advantage of using comparable corpora is that they are abundantly available, and can also be automatically extracted from the internet, covering recent topics [1].

In this paper, we propose a new method for automatic bilingual dictionary creation, using comparable corpora. Our method focuses on the extraction and comparison of *lexical-dependency* context across unrelated languages, like Japanese

# Online Learning via Dynamic Reranking for Computer Assisted Translation

Pascual Martínez-Gómez, Germán Sanchis-Trilles, and Francisco Casacuberta

Instituto Tecnológico de Informática  
Universidad Politécnica de Valencia  
{pmartinez, gsanchis, fcn}@dsic.upv.es

**Abstract.** New techniques for online adaptation in computer assisted translation are explored and compared to previously existing approaches. Under the online adaptation paradigm, the translation system needs to adapt itself to real-world changing scenarios, where training and tuning may only take place once, when the system is set-up for the first time. For this purpose, post-edit information, as described by a given quality measure, is used as valuable feedback within a dynamic reranking algorithm. Two possible approaches are presented and evaluated. The first one relies on the well-known perceptron algorithm, whereas the second one is a novel approach using the Ridge regression in order to compute the optimum scaling factors within a state-of-the-art SMT system. Experimental results show that such algorithms are able to improve translation quality by learning from the errors produced by the system on a sentence-by-sentence basis.

## 1 Introduction

Statistical Machine Translation (SMT) systems use mathematical models to describe the translation task and to estimate the probabilities involved in the process. [1] established the SMT grounds formulating the probability of translating a source sentence  $\mathbf{x}$  into a target sentence  $\hat{\mathbf{y}}$ , as

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \operatorname{Pr}(\mathbf{y} \mid \mathbf{x}) \quad (1)$$

In order to capture context information, *phrase-based* (PB) models [2,3] were introduced, widely outperforming single word models [4]. PB models were employed throughout this paper. The basic idea of PB translation is to segment the source sentence  $\mathbf{x}$  into *phrases* (i.e. word sequences), then to translate each source phrase  $\tilde{x}_k \in \mathbf{x}$  into a target phrase  $\tilde{y}_k$ , and finally reorder them to compose the target sentence  $\mathbf{y}$ .

Recently, the direct modelling of the posterior probability  $\operatorname{Pr}(\mathbf{y} \mid \mathbf{x})$  has been widely adopted. To this purpose, different authors [5,6] propose the use of the so-called log-linear models, where the decision rule is given by the expression

$$\begin{aligned} \hat{\mathbf{y}} &= \underset{\mathbf{y}}{\operatorname{argmax}} \sum_{m=1}^M \lambda_m h_m(\mathbf{x}, \mathbf{y}) \\ &= \underset{\mathbf{y}}{\operatorname{argmax}} \boldsymbol{\lambda} \cdot \mathbf{h}(\mathbf{x}, \mathbf{y}) = \underset{\mathbf{y}}{\operatorname{argmax}} s(\mathbf{x}, \mathbf{y}) \end{aligned} \quad (2)$$

where  $h_m(\mathbf{x}, \mathbf{y})$  is a score function representing an important feature for the translation of  $\mathbf{x}$  into  $\mathbf{y}$ ,  $M$  is the number of models (or features) and  $\lambda_m$  are the weights

# Learning Relation Extraction Grammars with Minimal Human Intervention: Strategy, Results, Insights and Plans

Hans Uszkoreit

German Research Center for Artificial Intelligence (DFKI) and Saarland University  
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany  
uszkoreit@dfki.de

**Abstract.** The paper describes the operation and evolution of a linguistically oriented framework for the minimally supervised learning of relation extraction grammars from textual data. Cornerstones of the approach are the acquisition of extraction rules from parsing results, the utilization of closed-world semantic seeds and a filtering of rules and instances by confidence estimation. By a systematic walk through the major challenges for this approach the obtained results and insights are summarized. Open problems are addressed and strategies for solving these are outlined.

**Keywords:** relation extraction, information extraction, minimally supervised learning, bootstrapping approaches to IE.

## 1 Introduction

While we still cannot build software systems that translate all or most sentences of a human language into some representation of their meanings, we are currently investigating methods for extracting relevant information from large volumes of texts. Some of the scientists working on information extraction view these methods as feasible substitutes for real text understanding, others see them as systematic steps toward a more comprehensive semantic interpretation. All agree on the commercial viability of effective information extraction applications, systems that detect references to interesting entities and to relevant relations between them, such as complex connections, properties, events and opinions.

One of the most intriguing but at the same time most challenging approaches to information extraction is the bootstrapping paradigm that starts from a very small set of semantic examples, called the *seed*, for discovering patterns or rules, which in turn are employed for finding additional instances of the targeted information type. These new instances will then be used as examples for the next round of finding linguistic patterns and the game repeats until no more instances can be detected. Since the seed can be rather small, containing between one and a handful of examples, this training scheme is usually called *minimally supervised learning*.

# Using Graph Based Method to Improve Bootstrapping Relation Extraction

Haibo Li, Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka

Graduate School of Information Science and Technology  
University of Tokyo

7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

lihaibo@mi.ci.i.u-tokyo.ac.jp, danushka@iba.t.u-tokyo.ac.jp,  
matsuo@biz-model.t.u-tokyo.ac.jp, ishizuka@i.u-tokyo.ac.jp

**Abstract.** Many bootstrapping relation extraction systems processing large corpus or working on the Web have been proposed in the literature. These systems usually return a large amount of extracted relationship instances as an out-of-ordered set. However, the returned result set often contains many irrelevant or weakly related instances. Ordering the extracted examples by their relevance to the given seeds is helpful to filter out irrelevant instances. Furthermore, ranking the extracted examples makes the selection of most similar instance easier. In this paper, we use a graph based method to rank the returned relation instances of a bootstrapping relation extraction system. We compare the used algorithm to the existing methods, relevant score based methods and frequency based methods, the results indicate that the proposed algorithm can improve the performance of the bootstrapping relation extraction systems.

## 1 Introduction

For many real world applications, background knowledge is intensively required. The acquisition of relational domain knowledge is still an important problem. Relation extraction systems extract structured relations from unstructured sources such as documents or web pages. These structured relations are as useful as knowledge. Acquiring relational facts *Acquirer–Acquiree* relation or *Person–Birthplace* relation with a small number of annotated data could have an important impact on applications such as business analysis research or automatic ontology construction.

Currently, research in relation extraction focuses mainly on pattern learning and matching techniques for extracting relational entity pairs from large corpora or the Web. The Web forms a fertile source of data for relation extraction, but users of relation extraction system are typically required to provide a large amount of annotated text to identify the interesting relation. This requirement is not feasible in real world applications. Therefore, many systems have been proposed to address the task of Web-based relation extraction, which usually only need a small number of seed entity pairs of relations. These systems typically build on the paradigm of bootstrapping of entity pairs and patterns as proposed by Brin[1].

However, an entity pair often has more than one type of semantic relations in real world. Consequently, a bootstrapping-based extraction system might introduce some

# A Hybrid Approach for the Extraction of Semantic Relations from MEDLINE Abstracts

Asma Ben Abacha and Pierre Zweigenbaum

LIMSI, CNRS, F-91403 Orsay, France  
{asma.benabacha,pz}@limsi.fr

**Abstract.** With the continuous digitisation of medical knowledge, information extraction tools become more and more important for practitioners of the medical domain. In this paper we tackle semantic relationships extraction from medical texts. We focus on the relations that may occur between diseases and treatments. We propose an approach relying on two different techniques to extract the target relations: (i) relation patterns based on human expertise and (ii) machine learning based on SVM classification. The presented approach takes advantage of the two techniques, relying more on manual patterns when few relation examples are available and more on feature values when a sufficient number of examples are available. Our approach obtains an overall 94.07% F-measure for the extraction of cure, prevent and side effect relations.

## 1 Introduction

Relation extraction is a long-standing research topic in Natural Language Processing, and has been used to help, among others, knowledge acquisition [1], information extraction [2], and question answering [3]. It has also received much attention in the medical [4] and biomedical domains [5]. With a large amount of information, health care professionals need fast and precise search tools such as question-answering systems [6]. Such systems need to correctly interpret (i) the questions and (ii) the texts from which answers will be extracted, hence the need for information extraction approaches such as [4,7,8]. The complexity of the task lies both in the linguistic issues known in open-domain tasks and in domain-specific features of the (bio)medical domain.

We propose here a hybrid approach to the detection of semantic relations in abstracts or full-text articles indexed by MEDLINE. This approach combines (i) a pattern-based method and (ii) a statistical learning method based on an SVM classifier which uses, among others, semantic resources. Their combination is based on a confidence score associated to the results of each method. We focus on extracting relations between a *disease* and a *treatment*. The obtained results are good and show the interest of combining both types of methods to disambiguate the multiple relations that can exist between two medical entities.

# An Active Learning Process for Extraction and Standardisation of Medical Measurements by a Trainable FSA

Jon Patrick and Mojtaba Sabbagh

Health Information Technology Research Laboratory, School of Information Technology,  
The University of Sydney, NSW, Australia  
jonpat@it.usyd.edu.au, mojtaba.sabbagh@sydney.edu.au

**Abstract.** Medical scores and measurements are a very important part of clinical notes as clinical staff infer a patient's state by analysing them, especially their variation over time. We have devised an active learning process for rapid training of an engine for detecting regular patterns of scores, measurements and people and places in clinical texts. There are two objectives to this task. Firstly, to find a comprehensive collection of validated patterns in a time efficient manner, and second to transform the captured examples into canonical forms. The first step of the process was to train an FSA from seed patterns and then use the FSA to extract further examples of patterns from the corpus.

The next step was to identify partial true positives (PTP) from the newly extracted examples. A manual annotator reviewed the extractions to identify the partial true positives (PTPs) and added the corrected form of these examples to the training set as new patterns. This cycle was continued until no new PTPs were detected. The process showed itself to be effective in requiring 5 cycles to create 371 true positives from 200 texts. We believe this gives 95% coverage of the TPs in the corpus.

**Keywords:** Finite State Automata, Medical Measurements, Active Learning.

## 1 Introduction

Our work specializes in processing corpora of medical texts [2][8]. These corpora usually contain many years of patient records from hospital clinical departments.

Clinical notes are a distinctly different genre of text with characteristics such as: 30% non-word tokens, idiosyncratic spellings, abbreviations and acronyms, and poor grammatical structure. The capacity to process the notes accurately is a direct function of learning something about them, e.g. the correct expansion of an abbreviation, and then reusing that immediately to understand subsequent text. In real-life situations the turnover of staff in a local hospital is so great that the language in use in the notes has its own dynamism that makes the continual accumulation of knowledge about the notes fundamental to a successful implementation of any practical medical language processing (MLP) technology that will survive the test of time.

# Topic Chains for Understanding a News Corpus

Dongwoo Kim and Alice Oh

KAIST  
Computer Science Department  
Daejeon, Korea  
dw.kim@kaist.ac.kr, alice.oh@kaist.edu

**Abstract.** The Web is a great resource and archive of news articles for the world. We present a framework, based on probabilistic topic modeling, for uncovering the meaningful structure and trends of important topics and issues hidden within the news archives on the Web. Central in the framework is a *topic chain*, a temporal organization of similar topics. We experimented with various topic similarity metrics and present our insights on how best to construct topic chains. We discuss how to interpret the topic chains to understand the news corpus by looking at long-term topics, temporary issues, and shifts of focus in the topic chains. We applied our framework to nine months of Korean Web news corpus and present our findings.

## 1 Introduction

The Web is a convenient and enormous source for learning about what is happening in the world. One can go to the Web site of any major news outlet or a portal site to get a quick overview of the important issues of the moment. However, it is difficult to use the Web to understand what has been happening over an extended period of time. We propose a computational framework based on probabilistic topic modeling to analyze a corpus of online news articles to produce results that show how the topics and issues emerge, evolve, and disappear within the corpus.

The problem of understanding a corpus of news articles over an extended period of time is challenging because one has to discover an unknown set of topics and issues from a large corpus of disparate sources, find and cluster similar topics, discover any short-term issues, and identify and display how the topics change over time. A narrower but similar problem has been studied in the TDT (topic detection and tracking) field [1] where the goal is to identify new events and track how they change over time. The events, however, are defined as happenings at certain places at certain times, and so they compose a small subset of general news topics and issues. For example, an earthquake in Haiti is an event, but the prolonged decline of real estate sales is not. The latter makes up a large portion of news, but the TDT community would only cover the former, whereas our research covers both. The probabilistic topic modeling community offers solutions such as Dynamic Topic Models [2] and Topics Over Time [3] for discovering topics

# From Italian Text to TimeML Document via Dependency Parsing

Livio Robaldo<sup>1</sup>, Tommaso Caselli<sup>2</sup>, Irene Russo<sup>2</sup>, and Matteo Grella<sup>3</sup>

<sup>1</sup> Department of Computer Science, University of Turin

<sup>2</sup> Istituto di Linguistica Computazionale, CNR, Pisa

<sup>3</sup> Parsit s.r.l.

<http://www.parsit.it/>

[robaldo@di.unito.it](mailto:robaldo@di.unito.it), [tommaso.caselli@ilc.cnr.it](mailto:tommaso.caselli@ilc.cnr.it),

[irene.russo@ilc.cnr.it](mailto:irene.russo@ilc.cnr.it), [matteo.grella@parsit.it](mailto:matteo.grella@parsit.it)

**Abstract.** This paper describes the first prototype for building TimeML xml documents starting from raw text for Italian. First, the text is parsed with the TULE parser, a dependency parser developed at the University of Turin. The parsed text is then used as input to the TimeML rule-based module we have implemented, henceforth called as ‘The converter’. So far, the converter identifies and classifies events in the sentence. The results are rather satisfactory, and this leads us to support the use of dependency syntactic relations for the development of higher level semantic tools.

## 1 Introduction

The access to information through content has become the new frontier in NLP. Innovative annotation schemes such as TimeML [12] have push forward this aspect by creating benchmark corpora. The TimeBank corpus [13] has renewed the interest in temporal processing and in its use for complex NLP task such as Open-Domain Question-Answering [16], Summarization and Information Extraction.

The task of temporal processing can be split into different subtasks. First, the basic ontological entities involved, i.e. events and temporal expressions, must be recognized and treated on their own. Then, temporal relations between them can be computed. This paper describes an implemented event detector and classifier, which represents the first step of an ongoing research collaboration on the development of a TimeML-compliant tool for Italian.

In TimeML, an event is defined as something that holds true, obtains/happens, or occurs. Natural language (NL) offers a variety of means to realize events, namely verbs, complex VPs (such as light verb constructions or idioms), nouns (including nominalizations, second order nominals and type-coercions), predicative constructions, prepositional phrases or adjectival phrases. Two innovative aspects introduced by TimeML with respect to event detection and classification are represented by:

# Self-adjusting Bootstrapping

Shoji Fujiwara<sup>1</sup> and Satoshi Sekine<sup>2</sup>

<sup>1</sup> Nikkei Digital Media, Inc.

shoji.fujiwara@nex.nikkei.co.jp

<sup>2</sup> Computer Science Department

New York University

sekine@cs.nyu.edu

**Abstract.** Bootstrapping has been used as a very efficient method to extract a group of items similar to a given set of seeds. However, the bootstrapping method intrinsically has several parameters whose optimal values differ from task to task, and from target to target. In this paper, first, we will demonstrate that this is really the case and serious problem. Then, we propose *self-adjusting bootstrapping*, where the original seed is segmented into the real seed and validation data. We initially bootstrap starting with the real seed, trying alternative parameter settings, and use the validation data to identify the optimal settings. This is done repeatedly with alternative segmentations in typical cross-validation fashion. Then the final bootstrapping is performed using the best parameter setting and the entire original seed set in order to create the final output. We conducted experiments to collect sets of company names in different categories. Self-adjusting bootstrapping substantially outperformed a baseline using a uniform parameter setting.

## 1 Introduction

Bootstrapping has been used in many information extraction tasks, such as harvesting names (Strzalkowski and Wang 96) (Collins and Singer 99), relations (Brin 98) (Agichtein and Gravano 00) (Ravichandran and Hovy 02) (Sun 09), and events (Yangarber et al. 00). Recently, there are more work on bootstrapping mostly using query logs (Pasca 07) (Pantel and Pennacchiotti 06) (Sekine and Suzuki 07). Given seeds of the desired names or relations (which we will hereafter call “items”), it gathers more items using a large un-annotated corpus. First, the most salient contexts of the seed items are found, then those contexts are used to find more items of the same kind. This process can be repeated to get more contexts and items. It is recognized as a very efficient method to extract a group of items similar to a given set of seeds, when there is enough data in the matrix of items and contexts. However, there is an essential problem in the bootstrapping method, namely parameter tuning. The bootstrapping method intrinsically has several parameters, such as the number of contexts to be used at each iteration, the number of items to be extracted at each iteration, and the scoring functions to calculate the similarity between contexts and between items. In

# Story Link Detection Based on Event Words

Letian Wang and Fang Li

Department of Computer Science & Engineering  
Shanghai Jiao Tong University  
Shanghai, China  
{koh, fli}@sjtu.edu.cn

**Abstract.** In this paper, we propose an event words based method for story link detection. Different from previous studies, we use time and places to label nouns and named entities, the featured nouns/named entities are called event words. In our approach, a document is represented by five dimensions including nouns/named entities, time featured nouns/named entities, place featured nouns/named entities, time&place featured nouns/named entities and publication date. Experimental results show that, our method gain a significant improvement over baseline and event words plays a vital role in this improvement. Especially when using publication date, we can reach the highest 92% on precision.

**Keywords:** story link detection, event words, multidimensional model, nouns/named entities, featured nouns/named entities.

## 1 Introduction

Story link detection, which was first defined in the Topic Detection and Tracking (TDT) [1,2,12,14,16] competition program, is the task of determining whether two stories, such as news articles and/or radio broadcasts, are about the same event, or linked. Story link detection is important for many applications. For example, there are three reports whose titles are:

- Midterm election polls open in United States
- US presidential vote is underway
- Voting in parliamentary election starts in Japan

The content of three news stories above are very similar, because they are all about the election, they have many common words in the text such as “election”, “vote”, “candidate” and so on. But actually they are different because they are not the same event. The first one is related to the election in U.S.A. in 2006 while the second one is about the election in U.S.A in 2008 and the last one refers to the election in Japan in 2007. The task of story link detection is to find out if the two stories are about the same event even though they may have the same content.

According to TDT, two stories are linked if the events in the stories happened at some specific time and place. In this paper, we give a more explicit definition:

# Ranking Multilingual Documents Using Minimal Language Dependent Resources

G.S.K. Santosh, N. Kiran Kumar, and Vasudeva Varma

International Institute of Information Technology, Hyderabad, India  
{santosh.gsk,kirankumar.n}@research.iiit.ac.in, vv@iiit.ac.in

**Abstract.** This paper proposes an approach of extracting simple and effective features that enhances multilingual document ranking (MLDR). There is limited prior research on capturing the concept of multilingual document similarity in determining the ranking of documents. However, the literature available has worked heavily with language specific tools, making them hard to reimplement for other languages. Our approach extracts various multilingual and monolingual similarity features using a basic language resource (bilingual dictionary). No language-specific tools are used, hence making this approach extensible for other languages. We used the datasets provided by Forum for Information Retrieval Evaluation (FIRE)<sup>1</sup> for their 2010 Adhoc Cross-Lingual document retrieval task on Indian languages. Experiments have been performed with different ranking algorithms and their results are compared. The results obtained showcase the effectiveness of the features considered in enhancing multilingual document ranking.

**Keywords:** Multilingual Document Ranking, Feature Engineering, Wikipedia, Levenshtein Edit Distance.

## 1 Introduction

Multilingual Information Retrieval (MLIR) is desirable with the increase of information in different languages. With the rapid development of globalization and digital online information in Internet, a growing demand for MLIR has emerged. MLIR involves the subtask of Cross Lingual Information Retrieval (CLIR) separately for each desired language. The clear separation of the retrieved result lists between different languages makes it necessary to have a merging step in order to produce a single result list. However, merging is intertwined with ranking step that ranks the documents of multilingual result lists as per the relevancy to the information need.

The problem of CLIR has been well studied in the past decade especially with the help of CLEF, NTCIR, TREC and FIRE forums. In the realm of CLIR the problem of ranking multilingual result lists is a very challenging task. The task of identifying whether two different language documents talks about the same

---

<sup>1</sup> <http://www.isical.ac.in/~clia/>

# Measuring Chinese-English Cross-Lingual Word Similarity with *HowNet* and Parallel Corpus

Yunqing Xia<sup>1</sup>, Taotao Zhao<sup>1,2</sup>, Jianmin Yao<sup>2</sup>, and Peng Jin<sup>3</sup>

<sup>1</sup> Department of Computer Science and Technology,  
Tsinghua University, Beijing 100084, China  
yqxia@tsinghua.edu.cn

<sup>2</sup> School of Computer Science and Technology,  
Soochow University, Suzhou 215006, China  
zhaott10@gmail.com, jyao@suda.edu.cn

<sup>3</sup> Lab of Intelligent Information Processing and Application,  
Leshan Normal University, Leshan 614004, China  
jandp@pku.edu.cn

**Abstract.** Cross-lingual word similarity (CLWS) is a basic component in cross-lingual information access systems. Designing a CLWS measure faces three challenges: (i) Cross-lingual knowledge base is rare; (ii) Cross-lingual corpora are limited; and (iii) No benchmark cross-lingual dataset is available for CLWS evaluation. This paper presents some Chinese-English CLWS measures that adopt *HowNet* as cross-lingual knowledge base and sentence-level parallel corpus as development data. In order to evaluate these measures, a Chinese-English cross-lingual benchmark dataset is compiled based on the Miller-Charles' dataset. Two conclusions are drawn from the experimental results. Firstly, *HowNet* is a promising knowledge base for the CLWS measure. Secondly, parallel corpus is promising to fine-tune the word similarity measures using cross-lingual co-occurrence statistics.

**Keywords:** Cross-lingual word similarity, cross-lingual information access, *HowNet*, parallel corpus.

## 1 Introduction

Word similarity plays a vital role in natural language processing and information retrieval. In natural language processing, word similarity is widely used in word sense disambiguation [1], synonym extraction [2]. In information retrieval, word similarity is adopted in multimodal documents retrieval [5] and image retrieval [6]. Human-compiled linguistic knowledge base (e.g., *WordNet* [7] and *HowNet* [8]) were widely used to measure word similarity [2,9-12]. As thesauri are usually static and incomplete, corpora were then adopted to estimate word similarity based on co-occurrence statistics [13-15]. Li et al. (2003) proved that thesaurus and corpus can be integrated to yield a better performance [16].

Cross-lingual word similarity (CLWS) reflects semantic similarity between two words in different languages. Very recently, CLWS research started to attract

# Comparing Manual Text Patterns and Machine Learning for Classification of E-Mails for Automatic Answering by a Government Agency

Hercules Dalianis<sup>1</sup>, Jonas Sjöbergh<sup>2</sup>, and Eriks Sneiders<sup>1</sup>

<sup>1</sup> Department of Computer and Systems Sciences (DSV),  
Stockholm University, Forum 100, SE-164 40 Kista, Sweden  
{eriks,hercules}@dsv.su.se

<sup>2</sup> KTH CSC,  
SE-100 44 Stockholm, Sweden  
jsh@kth.se

**Abstract.** E-mails to government institutions as well as to large companies may contain a large proportion of queries that can be answered in a uniform way. We analysed and manually annotated 4,404 e-mails from citizens to the Swedish Social Insurance Agency, and compared two methods for detecting answerable e-mails: manually-created text patterns (rule-based) and machine learning-based methods. We found that the text pattern-based method gave much higher precision at 89 percent than the machine learning-based method that gave only 63 percent precision. The recall was slightly higher (66 percent) for the machine learning-based methods than for the text patterns (47 percent). We also found that 23 percent of the total e-mail flow was processed by the automatic e-mail answering system.

**Keywords:** automatic e-mail answering, text pattern matching, machine learning, SVM, Naïve Bayes, E-government.

## 1 Introduction

Many governmental agencies and companies are today overwhelmed with e-mails with queries from citizens or customers that need an answer. Many of these e-mails are easy to reply to and do not need more advanced manual processing. The reply can even be made available on the web site of the government agency or the company. We studied the Swedish Social Insurance Agency (SSIA) (in Swedish “Försäkringskassan”<sup>1</sup>).

SSIA receives 350,000 e-mails per year, which are answered by 640 handling officers who also answer phone calls, use Internet chat, meet citizens and make decisions. The e-mail answering work in total corresponds to 25 full-time employees. If we could automatically answer even a fraction of these e-mails then much would be gained: citizens would obtain immediate answers and the workload of the handling officers would be reduced as they would not need to answer the most basic and monotonous e-mail queries and could focus on the more demanding ones and help citizens more effectively.

---

<sup>1</sup> <http://www.forsakringskassan.se>

# Using Thesaurus to Improve Multiclass Text Classification

Nooshin Maghsoodi and Mohammad Mehdi Homayounpour

Laboratory of Intelligent Signal and Speech Processing, Faculty of Computer Engineering,  
Amirkabir University of Technology, Tehran, Iran  
{n\_maghsoodi, homayoun}@aut.ac.ir

**Abstract.** With the growing amount of textual information available on the Internet, the importance of automatic text classification has been increasing in the last decade. In this paper, a system was presented for the classification of multiclass Farsi documents which uses Support Vector Machine (SVM) classifier. The new idea proposed in the present paper, is based on extending the feature vector by adding some words extracted from a thesaurus. The goal is to assist classifier when training dataset is not comprehensive for some categories. For corpus preparation, Farsi Wikipedia website and articles of some archived newspapers and magazines are used. As the results indicate, classification efficiency improves by applying this approach. 0.89 micro F-measure were achieved for classification of 10 categories of Farsi texts.

**Keywords:** Text classification, Support vector machine, Thesaurus, Farsi.

## 1 Introduction

The task of assigning natural language documents to a set of predefined categories is known as text classification. Due to the wide availability of online information in the World Wide Web, it may be impossible to classify the documents manually; so automatic classification of text documents seems to be inevitable. The workflow in most of the text classification systems is to train the classification system using a training dataset including many text documents whose categories are known. In the test phase, the system assigns a category to a new document. Each document in training dataset consists of a great number of relevant and irrelevant words corresponding to its category. One way to decrease the complexity of a text classifier and to increase its speed is to discard the irrelevant words and to render more weight on relevant ones. This phase which is called feature selection, is considered in many classification systems and different approaches such as the selection of features based on information gain, *tf\_idf* criterion and  $\chi^2$  test have been applied [1, 2, 3]. Classifier component in such systems is often one of the statistical methods or machine learning techniques including multivariate regression model, nearest neighbor classifier [3, 4], probabilistic Bayesian models [5, 6], decision tree [5, 7] and support vector machine (SVM) [8, 9, 10]. According studies carried out by Yang [10], SVM outperforms other machine learning methods. Therefore, in the approach presented in this paper, SVM is applied

# Adaptable Term Weighting Framework for Text Classification

Dat Huynh, Dat Tran, Wanli Ma, and Dharmendra Sharma

Faculty of Information Sciences and Engineering  
University of Canberra  
ACT 2601, Australia

{dat.huynh,dat.tran,wanli.ma,dharmendra.sharma}@canberra.edu.au

**Abstract.** In text classification, term frequency and term co-occurrence factors are dominantly used in weighting term features. Category relevance factors have recently been used to propose term weighting approaches. However, these approaches are mainly based on their own-designed text classifiers to adapt to category information, where the advantages of popular text classifiers have been ignored. This paper proposes a term weighting framework for text classification tasks. The framework firstly inherits the benefits of provided category information to estimate the weighting of features. Secondly, based on the feedback information, it is able to continuously adjust feature weightings to find the best representations for documents. Thirdly, the framework robustly makes it possible to work with different text classifiers on classifying the text representations, based on category information. On several corpora with SVM classifier, experiments show that given predicted information from TFxIDF method as initial status, the proposed approach leverages accuracy results and outperforms current text classification approaches.

**Keywords:** Text representation, feature weighting approach, term category dependency, classifier, and text classification.

## 1 Introduction

In text classification (TC), a single or multiple category labels are automatically assigned to a new text document based on category models, which are created after learning a set of labelled training text documents. TC methods normally convert a text document into a relational tuple using the popular vector-space model to obtain a list of terms with corresponding frequencies. A term-by-frequency matrix, interpreted as a relational table, is used to represent a collection of documents.

Due to the huge challenges and the difficulties of classifying document representations to a list of categories, a large number of classification algorithms have been developed to address the challenges in different degrees. Some of the popular algorithms have been currently used in TC such as multivariate regression

# Automatic Specialized vs. Non-specialized Sentence Differentiation

Iria da Cunha<sup>2,3,1</sup>, M. Teresa Cabré<sup>1</sup>, Eric SanJuan<sup>3</sup>,  
Gerardo Sierra<sup>2</sup>, Juan Manuel Torres-Moreno<sup>2,3,4</sup>, and Jorge Vivaldi<sup>1</sup>

<sup>1</sup> Institut Universitari de Linguística Aplicada - UPF  
Roc Boronat, 138 E-08018 Barcelona (Espanya)

<sup>2</sup> Grupo de Ingeniería Lingüística - Instituto de Ingeniería UNAM  
Torre de Ingeniería Aa Basamento, Ciudad Universitaria Mexico, D.F. 04510 Mexico

<sup>3</sup> Laboratoire Informatique d'Avignon - UAPV

339 chemin des Meinajaries, BP91228 84911 Avignon Cedex 9, France

<sup>4</sup> École Polytechnique de Montréal - Département de génie informatique  
CP 6079 Succ. Centre Ville H3C 3A7 Montréal (Québec), Canada

<http://www.lia.univ-avignon.fr>, <http://www.iula.upf.edu>,  
<http://www.iling.unam.mx>

**Abstract.** Compilation of Languages for Specific Purposes (LSP) corpora is a task which is fraught with several difficulties (mainly time and human effort), because it is not easy to discern between specialized and non-specialized text. The aim of this work is to study automatic specialized vs. non-specialized sentence differentiation. The experiments are carried out on two corpora of sentences extracted from specialized and non-specialized texts. One in economics (academic publications and news from newspapers), another about sexuality (academic publications and texts from forums and blogs). First we show the feasibility of the task using a statistical n-gram classifier. Then we show that grammatical features can also be used to classify sentences from the first corpus. For such purpose we use association rule mining.

**Keywords:** Specialized Text, General Text, Corpus, Languages for Specific Purposes, Statistical Methods, Association Rules, Grammatical features.

## 1 Introduction

Compilation of Languages for Specific Purposes (LSP) corpora, that is, corpora including specialized texts, is necessary to carry out several tasks, such as: terminology extraction, compiling specialized dictionaries, lexicons or ontologies. This corpora compilation is human time effort consuming. Until now, professionals or specialists have to decide if the text is specialized or not.

But what is a specialized text? [1] mentions some features to be considered in order to answer this question: the text author, the potential reader, the structural organization and the lexical units' selection. There are two types of variability in specialized texts: horizontal determined by the subject and vertical determined

# Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features

B. Thomas Adler<sup>1</sup>, Luca de Alfaro<sup>2</sup>, Santiago M. Mola-Velasco<sup>3</sup>,  
Paolo Rosso<sup>3</sup>, and Andrew G. West<sup>4,\*</sup>

<sup>1</sup> University of California, Santa Cruz, USA  
thumper@soe.ucsc.edu

<sup>2</sup> Google and UC Santa Cruz, USA  
luca@dealfaro.com

<sup>3</sup> NLE Lab. - ELiRF - DSIC. Universidad Politécnic de Valencia, Spain  
{smola,proso}@dsic.upv.es

<sup>4</sup> University of Pennsylvania, Philadelphia, USA  
westand@cis.upenn.edu

**Abstract.** Wikipedia is an online encyclopedia which anyone can edit. While most edits are constructive, about 7% are acts of vandalism. Such behavior is characterized by modifications made in bad faith; introducing spam and other inappropriate content.

In this work, we present the results of an effort to integrate three of the leading approaches to Wikipedia vandalism detection: a spatio-temporal analysis of metadata (STiki), a reputation-based system (WikiTrust), and natural language processing features. The performance of the resulting joint system improves the state-of-the-art from all previous methods and establishes a new baseline for Wikipedia vandalism detection. We examine in detail the contribution of the three approaches, both for the task of discovering fresh vandalism, and for the task of locating vandalism in the complete set of Wikipedia revisions.

## 1 Introduction

Wikipedia [1] is an online encyclopedia that anyone can edit. In the 10 years since its creation, 272 language editions have been created, with 240 editions being actively maintained as of this writing [2]. Wikipedia's English edition has more than 3 million articles, making it the biggest encyclopedia ever created. The encyclopedia has been a collaborative effort involving over 13 million registered users and an indefinite number of anonymous editors [2]. This success has made Wikipedia one of the most used knowledge resources available online and a source of information for many third-party applications.

The open-access model that is key to Wikipedia's success, however, can also be a source of problems. While most edits are constructive, some are *vandalism*,

---

\* Authors appear alphabetically. Order does not reflect contribution magnitude.

# Costco: Robust Content and Structure Constrained Clustering of Networked Documents

Su Yan<sup>1</sup>, Dongwon Lee<sup>2</sup>, and Alex Hai Wang<sup>3</sup>

<sup>1</sup> IBM Almaden Research Center  
San Jose, CA 95120, USA

<sup>2</sup> The Pennsylvania State University  
University Park, PA 16802, USA

<sup>3</sup> The Pennsylvania State University Dumore, PA 18512, USA  
syan@us.ibm.com,  
{dongwon,hwang}@psu.edu

**Abstract.** Connectivity analysis of networked documents provides high quality link structure information, which is usually lost upon a content-based learning system. It is well known that combining links and content has the potential to improve text analysis. However, exploiting link structure is non-trivial because links are often noisy and sparse. Besides, it is difficult to balance the term-based content analysis and the link-based structure analysis to reap the benefit of both. We introduce a novel networked document clustering technique that integrates the content and link information in a unified optimization framework. Under this framework, a novel dimensionality reduction method called COntent & STructure COnstrained (Costco) Feature Projection is developed. In order to extract robust link information from sparse and noisy link graphs, two link analysis methods are introduced. Experiments on benchmark data and diverse real-world text corpora validate the effectiveness of proposed methods.

**Keywords:** link analysis, dimensionality reduction, clustering.

## 1 Introduction

With the proliferation of the World Wide Web and Digital Libraries, analyzing “networked” documents has increasing challenge and opportunity. In addition to text content attributes, networked documents are correlated by links (e.g., hyperlinks between Web pages, citations between scientific publications etc.). These links are useful for text processing because they convey rich semantics that are usually independent of word statistics of documents [8].

Exploiting link information of networked documents to enhance text classification has been studied extensively in the research community [3,4,6,14]. It is found that, although both content attributes and links can independently form reasonable text classifiers, an algorithm that exploits both information sources has the potential to improve the classification [2,10]. Similar conclusion has been

# Learning Predicate Insertion Rules for Document Abstracting

Horacio Saggion

TALN

Department of Information and Communication Technologies  
Universitat Pompeu Fabra  
C/Tanger 122-140 Campus de la Comunicació Poble Nou - Barcelona  
08018 - Spain  
horacio.saggion@upf.edu

**Abstract.** The insertion of linguistic material into document sentences to create new sentences is a common activity in document abstracting. We investigate a transformation-based learning method to simulate this type of operation relevant for text summarization. Our work is framed on a theory of transformation-based abstracting where an initial text summary is transformed into an abstract by the application of a number of rules learnt from a corpus of examples. Our results are as good as recent work on classification-based predicate insertion.

## 1 Introduction

The problem of generating summaries by automatic means started in the early fifties [16] and continues nowadays to be a research topic receiving lot of attention [12,31,23,17,28,10]. The problem of generating “abstracts” – summaries containing linguistic material not necessarily present in the document to be summarized – has however received comparatively less attention. In this work we aim at simulating the way abstracts are produced and try to capture from textual data models of abstract production [26,11]. An example of the kind of abstract we aim to produce is shown in Figure 1. It is an abstract from the ERIC abstracting database which contains information extracted from the abstracted document together with rhetorical predicates inserted during abstract writing. These predicates inserted into the abstract by professional abstractors have specific communicative functions such as introducing the topic of the document, elaborating information, discussing particular issues, concluding, etc. used sometimes to improve the abstract and make it more objective [20]. Here we focus on this relatively new problem of combining document fragments with a limited set of linguistic expressions to create quasi-extractive summaries. The inserted predicates “glue” together the extracted fragments, thus creating a quasi-extractive summary. It is important to note that predicates can be prepended or appended to the sentence fragments, in the later case using a passive construction (e.g. “The state program in Rode Island is outlined”), note however, that we have

# Multi-topical Discussion Summarization Using Structured Lexical Chains and Cue Words

Jun Hatori<sup>1</sup>, Akiko Murakami<sup>2,3</sup>, and Jun'ichi Tsujii<sup>1,3,4,5</sup>

<sup>1</sup> Graduate School of Information Science and Technology, University of Tokyo  
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan  
{hatori, tsujii}@is.s.u-tokyo.ac.jp

<sup>2</sup> IBM Research – Tokyo  
1623-14 Shimotsuruma, Yamato, Kanagawa, Japan  
akikom@jp.ibm.com

<sup>3</sup> Graduate School of Interdisciplinary Information Studies, University of Tokyo  
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

<sup>4</sup> School of Computer Science, University of Manchester  
131 Princess Street, Manchester, M1 7DN, UK

<sup>5</sup> National Centre for Text Mining (NaCTeM), UK  
131 Princess Street, Manchester, M1 7DN, UK

**Abstract.** We propose a method to summarize threaded, multi-topical texts automatically, particularly online discussions and e-mail conversations. These corpora have a so-called reply-to structure among the posts, where multiple topics are discussed simultaneously with a certain level of continuity, although each post is typically short. We specifically focus on the multi-topical aspect of the corpora, and propose the use of two linguistically motivated features: lexical chains and cue words, which capture the topics and topic structure. Particularly, we introduce the *structured lexical chain*, which is a combination of traditional lexical chains with the thread structure. In experiments, we show the effectiveness of these features on the Innovation Jam 2008 Corpus and the BC3 Mailing List Corpus based on two task settings: key-sentence and keyword extraction. We also present detailed analysis of the result with some intuitive examples.

## 1 Introduction

Online discussion has become a popular tool for collaboration among people as they discuss various topics online. However, with its increasing popularity, problems have arisen with information overload, which makes it difficult for people to catch up with up-to-date topics and central points of the discussion. Particularly, if organizers intend to draw out useful findings from the whole discussion, they often encounter a problem with obtaining the big picture of the content that is distributed among a large number of posts. Therefore, great demand exists for systems that provide users with an overview of the discussion.

Posts in an online discussion are typically organized in either a sequential or a tree-structured thread. Although the former has simpler structure, the latter allows division of many topics into smaller branches. For this reason, the tree-structured thread has been

# Multi-document Summarization Using Link Analysis Based on Rhetorical Relations between Sentences

Nik Adilah Hanin Binti Zahri and Fumiyo Fukumoto

Interdisciplinary Graduate School of Medicine and Engineering  
University of Yamanashi, Japan  
{g09dh103, fukumoto}@yamanashi.ac.jp

**Abstract.** With the accelerating rate of data growth on the Internet, automatic multi-document summarization has become an important task. In this paper, we propose a link analysis incorporated with rhetorical relations between sentences to perform extractive summarization for multiple-documents. We make use of the documents headlines to extract sentences with salient terms from the documents set using statistical model. Then we assign rhetorical relations learned by SVMs to determine the connectivity between the sentences which include the salient terms. Finally, we rank these sentences by measuring their relative importance within the document set based on link analysis method, PageRank. The rhetorical relations are used to evaluate the complementarity and redundancy of the ranked sentences. Our evaluation results show that the combination of PageRank along with rhetorical relations among sentences does help to improve the quality of extractive summarization.

**Keywords:** Probability model, n-gram, link-based analysis, Support Vector Machine, extractive summarization, rhetorical relations.

## 1 Introduction

Due to rapid growth of information on the Internet recently, finding specific data from huge amount of document is crucial since it requires a lot of time and efforts for users to read each document. As a result, automatic summarization has become an important technique nowadays. Text summarization helps to simplify information search and cut the search time by pointing the most relevant information which allows users to quickly comprehend the information contained in a large document.

The general approach of automatic text summarization is extractive or abstractive summarization. Extractive summarization focuses in finding the most salient sentences from the original document, while abstractive summarization focuses on generating summary by selecting only important terms from documents and might not contain original phrase or word. Our work focuses on extractive summarization. Previous works in this area have proposed various techniques such as, centroid-based summarization method [1], automated document indexing based on statistical latent model [2] and most recent technique, text summarization based on Cross-document Structure Theory (CST) relationship between sentences[3][4][5].

# Co-clustering Sentences and Terms for Multi-document Summarization

Yunqing Xia<sup>1</sup>, Yonggang Zhang<sup>1,2</sup>, and Jianmin Yao<sup>2</sup>

<sup>1</sup> Department of Computer Science and Technology,  
Tsinghua University, Beijing 100084, China  
yqxia@tsinghua.edu.cn

<sup>2</sup> School of Computer Science and Technology,  
Soochow University, Suzhou 215006, China  
yonggang118@gmail.com, jyao@suda.edu.cn

**Abstract.** Two issues are crucial to multi-document summarization: diversity and redundancy. Content within some topically-related articles are usually redundant while the topic is delivered from diverse perspectives. This paper presents a co-clustering based multi-document summarization method that makes full use of the diverse and redundant content. A multi-document summary is generated in three steps. First, the sentence-term co-occurrence matrix is designed to reflect diversity and redundancy. Second, the co-clustering algorithm is performed on the matrix to find globally optimal clusters for sentences and terms in an iterative manner. Third, a more accurate summary is generated by selecting representative sentences from the optimal clusters. Experiments on DUC2004 dataset show that the co-clustering based multi-document summarization method is promising.

**Keywords:** Co-clustering, multi-document summarization, term extraction.

## 1 Introduction

Handling a large set of topically-related articles manually is usually laborious and time-consuming. Aiming at generating a summary that covers the major themes in an article collection, multi-document summarization provides a promising solution to the information overload problem. An ideal multi-document summary should cover not only the key topic of the multi-documents but also the diverse views of the multi-documents. Two distinct characteristics make multi-document summarization rather different from single-document summarization: diversity and redundancy [1-5]. Content within some topically-related articles are usually redundant while the topic is delivered from diverse perspectives. This is because the writers usually show common interests on popular target but they tend to report the target from different perspectives. As a result, diversity among the articles tends to be significant. However, some background information is usually necessary for the readers to follow the story. Therefore, redundant sentences are constantly found within the articles.

# Answer Validation Using Textual Entailment

Partha Pakray<sup>1</sup>, Alexander Gelbukh<sup>2</sup>, and Sivaji Bandyopadhyay<sup>1</sup>

<sup>1</sup> Computer Science and Engineering Department,  
Jadavpur University, Kolkata, India

<sup>2</sup> Center for Computing Research, National Polytechnic Institute,  
Mexico City, Mexico

parthapakray@gmail.com, www.gelbukh.com,  
sbandyopadhyay@cse.jdvu.ac.in

**Abstract.** We present an Answer Validation System (AV) based on Textual Entailment and Question Answering. The important features used to develop the AV system are Lexical Textual Entailment, Named Entity Recognition, Question-Answer type analysis, chunk boundary module and syntactic similarity module. The proposed AV system is rule based. We first combine the question and the answer into Hypothesis (H) and the Supporting Text as Text (T) to identify the entailment relation as either “VALIDATED” or “REJECTED”. The important features used for the lexical Textual Entailment module in the present system are: WordNet based unigram match, bigram match and skip-gram. In the syntactic similarity module, the important features used are: subject-subject comparison, subject-verb comparison, object-verb comparison and cross subject-verb comparison. The results obtained from the answer validation modules are integrated using a voting technique. For training purpose, we used the AVE 2008 development set. Evaluation scores obtained on the AVE 2008 test set show 66% precision and 65% F-Score for “VALIDATED” decision.

**Keywords:** Answer Validation Exercise (AVE), Textual Entailment (TE), Named Entity (NE), Chunk Boundary, Syntactic Similarity, Question Type.

## 1 Introduction

Answer Validation Exercise (AVE) is a task introduced in the QA@CLEF competition. AVE task is aimed at developing systems that decide whether the answer of a Question Answering system is correct or not. There were three AVE competitions AVE 2006 [1], AVE 2007 [2] and AVE 2008 [3]. AVE systems receive a set of triplets (Question, Answer and Supporting Text) and return a judgment of “SELECTED”, “VALIDATED” or “REJECTED” for each triplet. The evaluation methodology was improved over the years and oriented to identify the useful factors for QA systems improvement. Thus, in 2007 the AVE systems were to select only one VALID answer for every question from a set of possible answers, whereas in 2006, several VALID answers were possible to be selected. In 2008<sup>1</sup>, the organizers

---

<sup>1</sup> <http://nlp.uned.es/clef-qa/ave/>

# SPIDER: A System for Paraphrasing in Document Editing and Revision — Applicability in Machine Translation Pre-editing

Anabela Barreiro

Centro de Linguística da Universidade do Porto, Portugal  
barreiro\_anabela@hotmail.com

**Abstract.** This paper presents SPIDER, a system for paraphrasing in document editing and revision with applicability in machine translation pre-editing. SPIDER applies its linguistic knowledge (dictionaries and grammars) to create paraphrases of distinct linguistic phenomena. The first version of this tool was initially developed for Portuguese (ReEscreve v01), but it is extensible to different languages and can also operate across languages. SPIDER has a totally new interface, new resources which contemplate a wider coverage of linguistic phenomena, and applicability to legal terminology, which is described here.

**Keywords:** paraphrase, language composition tool, authoring aid, text processing application, pre-editing, revision, linguistic quality assurance.

## 1 Introduction

The relevance of paraphrases for natural language processing has been clearly defined, and paraphrases are being used in different types of applications for a variety of purposes. Paraphrasal knowledge plays a very important role in interpretation and generation of natural language. In *natural language interpretation*, dynamic semantics and identical parses resulting from paraphrases are important to successful applications. In *natural language generation*, the generation of paraphrases allows more varied and fluent text to be produced [Iordanskaja et al. 1991]. In *multi-document summarization*, the identification of paraphrases allows information across documents to be condensed [McKeown et al. 2002] and helps improve the quality of the generated summaries [Hirao et al. 2004]. In *question answering*, discovering paraphrased answers may provide additional evidence that an answer is correct [Ibrahim et al. 2003]. Paraphrases can also be useful in *text mining*, preventing a passage being discarded due to the inability to match a question phrase deemed as very important [Paşca and Dienes 2005]. In *information extraction*, paraphrases help text categorization tasks or mapping to texts with similar characteristics, lessening the disparity in the trigger word or the applicable extraction pattern [Shinyama and Sekine 2005]. Paraphrasing also helps *machine translation* performance [Callison-Burch 2007], in particular the translation of multi-word units [Barreiro 2011].

# Providing Cross-Lingual Editing Assistance to Wikipedia Editors

Ching-man Au Yeung, Kevin Duh, and Masaaki Nagata

NTT Communication Science Laboratories  
2-4 Hikaridai, Seika-cho, Soraku-gun  
Kyoto, 619-0237, Japan

{[auyeung,kevinduh](mailto:auyeung,kevinduh}@cslab.kecl.ntt.co.jp)}@cslab.kecl.ntt.co.jp, [nagata.masaaki@lab.ntt.co.jp](mailto:nagata.masaaki@lab.ntt.co.jp)

**Abstract.** We propose a framework to assist Wikipedia editors to transfer information among different languages. Firstly, with the help of some machine translation tools, we analyse the texts in two different language editions of an article and identify information that is only available in one edition. Next, we propose an algorithm to look for the most probable position in the other edition where the new information can be inserted. We show that our method can accurately suggest positions for new information. Our proposal is beneficial to both readers and editors of Wikipedia, and can be easily generalised and applied to other multi-lingual corpora.

## 1 Introduction

There are currently over 250 different language editions in Wikipedia. However, significant differences exist between different editions in terms of size and quality [6]. Several projects on Wikipedia have been initiated to bridge this information gap with the help of both human and machine translation [12,13,14]. Google also provides a translator toolkit that assists users to translate Wikipedia articles<sup>1</sup>.

While existing efforts focused on translating whole articles, we believe maintaining existing articles across different languages is also a major challenge. Wikipedia is by no means a static encyclopedia. Articles are constantly being revised by editors. As different language editions are being developed separately, it is likely that different language editions will contain different information, depending on the focuses of the editors or interests of the respective community.

Although Wikipedia is not intended to be an encyclopedia in which different language editions are exact translations of one another [14], it is desirable to keep any article up-to-date and comprehensive. However, the effort required to identify what should be translated can be prohibitively expensive, especially when the target document already has substantial content. This requires editors to continuously monitor articles in different languages, which is clearly unscalable.

We propose a framework that assists Wikipedia editors or translators to transfer information from one language into another. We term this task **cross-lingual document enrichment**. Our proposed framework is completely automatic and

---

<sup>1</sup> Google Translator Toolkit: <http://translate.google.com/toolkit>

# Reducing Overdetections in a French Symbolic Grammar Checker by Classification

Fabrizio Gotti<sup>1</sup>, Philippe Langlais<sup>1</sup>, Guy Lapalme<sup>1</sup>,  
Simon Charest<sup>2</sup>, and Éric Brunelle<sup>2</sup>

<sup>1</sup> DIRO/Univ. de Montréal  
C.P. 6128, Succ Centre-Ville  
H3C 3J7 Montréal (Québec) Canada  
<http://rali.iro.umontreal.ca>

<sup>2</sup> Druide Informatique  
1435 rue Saint-Alexandre, bureau 1040  
H3A 2G4 Montréal (Québec) Canada  
<http://www.druide.com>

**Abstract.** We describe the development of an “overdetection” identifier, a system for filtering detections erroneously flagged by a grammar checker. Various families of classifiers have been trained in a supervised way for 14 types of detections made by a commercial French grammar checker. Eight of these were integrated in the most recent commercial version of the system. This is a striking illustration of how a machine learning component can be successfully embedded in *Antidote*, a robust, commercial, as well as popular natural language application.

## 1 Introduction

Even though most modern writers use, often unknowingly, the grammar checker embedded in Microsoft Word, few NLP researchers have addressed the problem of improving the quality of the grammatical error detection algorithms [1,2]. Clément et al. [3] suggest that this could be explained by the lack of an annotated error corpus and by the close link that exists between a grammar checker and the proprietary word processor that embeds it.

Bustamante and Léon [4] present a typology of errors often encountered in Spanish and describe how the GramCheck project dealt with them. They distinguish structural errors (e.g. bad prepositional attachments) from non structural ones (e.g. subject verb agreement). The former is dealt with by crafting rules encoding typical errors that are added to the language parsing rules or by using auxiliary grammars on an ad hoc basis. The latter is dealt by loosening the unification process within the parser. These developments are quite complex and require a fine tuning of linguistic heuristics used within the parsing process.

Two main approaches to grammar checking have been taken by researchers. The first approach consists in comparing the sentence to proofread against a model of proper language use (a positive grammar). For instance, [5] propose using  $n$ -grams to create a language model of lemmas and part-of-speech tags (POS) occurring in proper English text. The second strategy seeks to create

# Performance Evaluation of a Novel Technique for Word Order Errors Correction Applied to Non Native English Speakers' Corpus

Theologos Athanaselis, Stelios Bakamidis, and Ioannis Dologlou

Institute for Language and Speech Processing (ILSP) / R.C "ATHENA"

Artemidos 6 and Epidavrou, GR-15125,

Athens, Greece

{tathana,bakam,ydol}@ilsp.gr

**Abstract.** This work presents the evaluation results of a novel technique for word order errors correction, using non native English speakers' corpus. This technique, which is language independent, repairs word order errors in sentences using the probabilities of most typical trigrams and bigrams extracted from a large text corpus such as the British National Corpus (BNC). A good indicator of whether a person really knows a language is the ability to use the appropriate words in a sentence in correct word order. The "scrambled" words in a sentence produce a meaningless sentence. Most languages have a fairly fixed word order. For non-native speakers and writers, word order errors are more frequent in English as a Second Language. These errors come from the student if he is translating (thinking in his/her native language and trying to translate it into English). For this reason, the experimentation task involves a test set of 50 sentences translated from Greek to English. The purpose of this experiment is to determine how the system performs on real data, produced by non native English speakers.

**Keywords:** Word order errors; statistical language model; permutations filtering; British National Corpus; non native English speakers' corpus.

## 1 Introduction

Research on detecting erroneous sentences can be mainly classified into three categories. The first category makes use of hand-crafted rules [1],[2],[3]. These methods have been shown to be effective in detecting certain kinds of grammatical errors, but it is expensive to write non-conflicting rules in order to cover the wide range of grammatical errors. The second category focuses on parsing ill-formed sentences [4],[5],[6],[7]. The third category uses statistical techniques to detect erroneous sentences. Instead of asking experts to write hand-crafted rules, statistical approaches [8],[9],[10],[11] build statistical models to indentify sentences containing errors.

There are also other studies on detecting grammar errors at sentence level. More [12] introduced an English grammar checker for non-native English speakers. Heift

# Correcting Verb Selection Errors for ESL with the Perceptron

Xiaohua Liu<sup>1,3</sup>, Bo Han<sup>2,\*</sup>, and Ming Zhou<sup>3</sup>

<sup>1</sup> School of Computer Science and Technology,  
Harbin Institute of Technology, Harbin, 150001, China  
xiaoliu@microsoft.com

<sup>2</sup> Department of Computer Science and Software Engineering,  
The University of Melbourne, Victoria 3010, Australia  
b.han@pgrad.unimelb.edu.au

<sup>3</sup> Microsoft Research Asia, Beijing, 100190, China  
mingzhou@microsoft.com

**Abstract.** We study the task of correcting verb selection errors for English as a Second Language (ESL) learners, which is meaningful but also challenging. The difficulties of this task lie in two aspects: the lack of annotated data and the diversity of verb usage context. We propose a perceptron based novel approach to this task. More specifically, our method generates correction candidates using predefined confusion sets, to avoid the tedious and prohibitively unaffordable human labeling; moreover, rich linguistic features are integrated to represent verb usage context, using a global linear model learnt by the perceptron algorithm. The features used in our method include a language model, local text, chunks, and semantic collocations. Our method is evaluated on both synthetic and real-world corpora, and consistently achieves encouraging results, outperforming all baselines.

**Keywords:** verb selection, perceptron learning, ESL.

## 1 Introduction

Learners of English as a second language (ESL) are a large and growing section of the world's population. They tend to make various errors in English writing, among which, verb selection errors can be quite confusing and misleading. For example, in the sentence (written by a Chinese), “*I often **play with** my friends at school*”, the intended meaning of “*to play with*” is “*to have fun with one's friends*”. However, “*play with*” in English is often understood as “*to play (a game) with*”, “*to play with (among young children)*”, or “*to treat somebody or something frivolously*”; thus it deviates in a subtle way from the meaning intended by the Chinese speaker.

Therefore, designing such a device that can automatically detect and correct verb selection errors made by ESL learners is meaningful. However, this task is

---

\* This work has been done while the author was visiting Microsoft Research Asia.

# A Posteriori Agreement as a Quality Measure for Readability Prediction Systems

Philip van Oosten<sup>1,2</sup>, Véronique Hoste<sup>1,2</sup>, and Dries Tanghe<sup>1,2</sup>

<sup>1</sup> LT<sup>3</sup> Language and Translation Technology Team, University College Ghent,  
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium

<sup>2</sup> Ghent University, Krijgslaan 281, 9000 Ghent, Belgium

**Abstract.** All readability research is ultimately concerned with the research question whether it is possible for a prediction system to automatically determine the level of readability of an unseen text. A significant problem for such a system is that readability might depend in part on the reader. If different readers assess the readability of texts in fundamentally different ways, there is insufficient a priori agreement to justify the correctness of a readability prediction system based on the texts assessed by those readers. We built a data set of readability assessments by expert readers. We clustered the experts into groups with greater a priori agreement and then measured for each group whether classifiers trained only on data from this group exhibited a classification bias. As this was found to be the case, the classification mechanism cannot be unproblematically generalized to a different user group.

## 1 Introduction

In the most general terms, the goal of authoring a text is to get a message across to an intended audience. The readability of a text, then, can be defined as the relative ease of that audience to understand the author's message. It is intuitively clear that, even when defined in such general terms, the inherent subjectivity of the concept of readability cannot be ignored. The ease with which a given reader can correctly identify the message conveyed in a text is, among other things, inextricably related to the reader's background knowledge of the subject at hand [11].

The domain of readability research has at its primary research goal the design of a method to automatically predict the readability of a text. In recent years, a tendency seems to have arisen to explicitly address the subjective aspect of readability. [14] ultimately base their readability prediction method exclusively on the extent to which readers found a text to be "well-written". [10] take the assessments supplied by a number of experts as their gold standard, and test their readability prediction method as well as assessments by novices against these expert opinions. Similarly, [13] compile a gold standard for readability prediction by collecting assessments by expert and naive readers.

Subjective assessment entails the problem of reliably aggregating data that were obtained from various sources. This is a recurring issue in Natural Language

# A Method to Measure the Reading Difficulty of Japanese Words

Keiji Yasuda, Andrew Finch, and Eiichiro Sumita

National Institute of Information and Communications Technology  
3-5, Hikaridai, Keihanna Science City, Kyoto, 619-0289, Japan  
{keiji.yasuda, andrew.finch, eiichiro.sumita}@nict.go.jp

**Abstract.** In this paper, we propose an automatic method to measure the reading difficulty of Japanese words. The proposed method uses a statistical transliteration framework, which was inspired by statistical machine translation research. A Dirichlet process model is used for the alignment between single kanji characters and one or more hiragana characters. The joint probability of kanji and hiragana is used to measure the difficulty. In our experiment, we carried out a linear discriminate analysis using three kinds of lexicons: a Japanese place name lexicon, a Japanese last name lexicon and a general noun lexicon. We compared the discrimination ratio given by the proposed method and the conventional method, which estimates a word difficulty based on manually defined kanji difficulty. According to the experimental results, the proposed method performs well for scoring Japanese proper noun reading difficulty. The proposed method produces a higher discrimination ratio with the proper noun lexicons (14 points higher on the place name lexicon and 26.5 points higher on the last name lexicon) than the conventional method.

## 1 Introduction

The Japanese writing system uses two sets of phonograms (hiragana and katakana) and one set of logograms (“kanji,” or Chinese characters). A single kanji has one or more possible readings, which are categorized as “onyomi” (Sino-Japanese reading) or “kunyomi” (Japanese reading). In some cases, a single kanji character can have more than 10 different readings.

A Japanese single word is written using one or more kanji characters. Depending on the anteroposterior characters or even on context, the reading of the kanji can change. Consequently, some Japanese words are very difficult to read even for native Japanese speakers. In this paper, we propose a method to measure the reading difficulty of Japanese words, which could have practical application in language learning.

The difficulty of Japanese word readings can be attributed to two factors. The first factor is the difficulty of the kanji. There are 50,000 kanji characters in total. Only 2136 characters have been designated for everyday use, and rarely used kanji characters can be difficult to read. The second factor is the irregularity

# Informality Judgment at Sentence Level and Experiments with Formality Score

Shibamouli Lahiri, Prasenjit Mitra, and Xiaofei Lu

The Pennsylvania State University, University Park PA 16802, USA  
shibamouli@cse.psu.edu, pmitra@ist.psu.edu, xx113@psu.edu

**Abstract.** Formality and its converse, informality, are important dimensions of authorial style that serve to determine the social background a particular document is coming from, and the potential audience it is targeted to. In this paper we explored the concept of formality at the sentence level from two different perspectives. One was the Formality Score (F-score) and its distribution across different datasets, how they compared with each other and how F-score could be linked to human-annotated sentences. The other was to measure the inherent agreement between two independent judges on a sentence annotation task. It gave us an idea how subjective the concept of formality was at the sentence level. Finally, we looked into the related issue of document readability and measured its correlation with document formality.

## 1 Introduction

Writing style is an important dimension of human languages. Two documents can provide the same content, but they may have been written using very different styles [9]. Authors from different social, educational and cultural backgrounds tend to use different writing styles [4]. With the evolution of Web 2.0, user-generated content has given rise to a variety of writing styles. Blog posts, for example, are written differently from the way academic papers are written. Twitter chats manifest yet another kind of writing style. Wikipedia articles use their own style guide<sup>1</sup>.

One prominent dimension of writing style is the formality of a document. Academic papers are usually considered more formal than online forum posts. The notions of formality and contextuality at the document level have been illustrated by Heylighen and Dewaele [7]. They proposed a frequentist statistic known as the Formality Score (F-score) of a document, based on the number of deictic and non-deictic words (cf. Section 2). F-score is a coarse-grain measure, but it works well when used to classify documents according to their authorial style [15].

Classifying sub-document units such as sentences as formal or informal is more difficult because they are typically much smaller than a document and provide much less information. For example, the sentence “She doesn’t like the piano” may be considered informal because it contains the colloquial usage “doesn’t”. But some native English speakers may think that the usage of “doesn’t” is quite appropriate and formal. So we note that the notion of formality at the sentence level is subjective. On the other

---

<sup>1</sup> [http://en.wikipedia.org/wiki/Wikipedia:Manual\\_of\\_Style](http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style)

# Combining Word and Phonetic-Code Representations for Spoken Document Retrieval

Alejandro Reyes-Barragán<sup>1</sup>, Manuel Montes-y-Gómez<sup>1,2</sup>,  
and Luis Villaseñor-Pineda<sup>1</sup>

<sup>1</sup> Laboratory of Language Technologies,  
National Institute of Astrophysics, Optics and Electronics (INAOE),  
Luis Enrique Erro #1, Sta. María Tonantzintla, Puebla, Mexico  
{alejandroreyes, mmontesg, villasen}@inaoep.mx

<sup>2</sup> Department of Computer and Information Sciences,  
The University of Alabama at Birmingham (UAB),  
1300 University Boulevard, Birmingham, Alabama, USA

**Abstract.** The traditional approach for spoken document retrieval (SDR) uses an automatic speech recognizer (ASR) in combination with a word-based information retrieval method. This approach has only showed limited accuracy, partially because ASR systems tend to produce transcriptions of spontaneous speech with significant word error rate. In order to overcome such limitation we propose a method which uses word and phonetic-code representations in collaboration. The idea of this combination is to reduce the impact of transcription errors in the processing of some (presumably complex) queries by representing words with similar pronunciations through the same phonetic code. Experimental results on the CLEF-CLSR-2007 corpus are encouraging; the proposed hybrid method improved the mean average precision and the number of retrieved relevant documents from the traditional word-based approach by 3% and 7% respectively.

## 1 Introduction

The large amount of information existing in spoken form, such as TV and radio broadcasts, recordings of meetings, lectures and telephone conversations, has motivated the development of new technologies for its searching and browsing. Particularly, spoken document retrieval (SDR) refers to the task of finding segments from recorded speech that are relevant to a user's information need [1].

The traditional approach for SDR consists in a simple concatenation of an automatic speech recognition (ASR) system with a standard word-based retrieval method [2]. The main inconvenience of this approach is that it greatly depends on the accuracy of the recognition output. It is well known that recognition errors usually degrade the effectiveness of a SDR system, and that, unfortunately, current ASR methods have word error rates that vary from 20% to 40% in accordance to the kind of discourse.

# Automatic Rule Extraction for Modeling Pronunciation Variation\*

Zeeshan Ahmed and Julie Carson-Berndsen

CNGL, School of Computer Science and Informatics  
University College Dublin, Ireland  
zeeshan.ahmed@ucdconnect.ie, julie.berndsen@ucd.ie

**Abstract.** This paper describes the technique for automatic extraction of pronunciation rules from continuous speech corpus. The purpose of the work is to model pronunciation variation in phoneme based continuous speech recognition at language model level. In modeling pronunciation variations, morphological variations and out-of-vocabulary words problem are also implicitly modeled in the system. It is not possible to model these kind of variations using dictionary based approach in phoneme based automatic speech recognition. The variations are automatically learned from annotated continuous speech corpus. The corpus is first aligned, on the basis of phoneme and letter, using a dynamic string alignment algorithm. The DSA is applied to isolated words to deal with intra-word variations as well as to complete sentences in the corpus to deal with inter-word variations. The pronunciation rules *phonemes*  $\rightarrow$  *letters* are extracted from these aligned speech units to build pronunciation model. The rules are finally fed to a phoneme-to-word decoder for recognition of the words having different pronunciations or that are OOV.

## 1 Introduction

Pronunciation variations and treatment of out-of-vocabulary (OOV) words in automatic speech recognition (ASR) have always emerged as one of the biggest drawbacks for an ASR system. Pronunciation variation can occur because of dialect, native and non-native speaker, age, gender, emotions, position of words in the utterance etc.

Pronunciation variations can be incorporated at different levels in ASR system as explained in [1]. There are three levels at which pronunciation variations can be modeled: the lexicon, the acoustic model, the language model. To deal with pronunciation variations at the lexicon level, different variants of word pronunciation are added to the lexicon. At the acoustic level, context dependent phone modeling [2,3] has been widely used to capture the phone variations

---

\* This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Center for Next Generation Localization ([www.cngl.ie](http://www.cngl.ie)) at University College Dublin. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

# Predicting Word Pronunciation in Japanese

Jun Hatori<sup>1,\*</sup> and Hisami Suzuki<sup>2</sup>

<sup>1</sup> Department of Computer Science, University of Tokyo  
7-3-1 Hongo, Bunkyo, Tokyo, 113-0033 Japan  
hatori@is.s.u-tokyo.ac.jp

<sup>2</sup> Microsoft Research  
One Microsoft Way, Redmond, WA 98052, USA  
hisamis@microsoft.com

**Abstract.** This paper addresses the problem of predicting the pronunciation of Japanese words, especially those that are newly created and therefore not in the dictionary. This is an important task for many applications including text-to-speech and text input method, and is also challenging, because Japanese kanji (ideographic) characters typically have multiple possible pronunciations. We approach this problem by considering it as a simplified machine translation/transliteration task, and propose a solution that takes advantage of the recent technologies developed for machine translation and transliteration research. More specifically, we divide the problem into two subtasks: (1) Discovering the pronunciation of new words or those words that are difficult to pronounce by mining unannotated text, much like the creation of a bilingual dictionary using the web; (2) Building a decoder for the task of pronunciation prediction, for which we apply the state-of-the-art discriminative substring-based approach. Our experimental results show that our classifier for validating the word-pronunciation pairs harvested from unannotated text achieves over 98% precision and recall. On the pronunciation prediction task of unseen words, our decoder achieves over 70% accuracy, which significantly improves over the previously proposed models.

**Keywords:** Japanese language, pronunciation prediction, substring-based transliteration, letter-to-phoneme.

## 1 Introduction

This paper explores the problem of assigning pronunciation to words, especially when they are new and therefore not in the dictionary. The task is naturally important for the text-to-speech application [27], and has been researched in that context as letter-to-phoneme conversion, which converts an orthographic character sequence into phonemes. In addition to speech applications, the task is also crucial for those languages that require pronunciation-to-character conversion to input text, such as Chinese and Japanese, where users generally type in the pronunciations of words,

---

\* This work was conducted during the first author's internship at Microsoft Research.

# A Minimum Cluster-Based Trigram Statistical Model for Thai Syllabification

Chonlasith Jucksriporn and Ohm Sornil

Department of Computer Science  
National Institute of Development Administration  
Bangkok 10240, Thailand  
chonlasith@gmail.com, osornil@as.nida.ac.th

**Abstract.** Syllabification is a process of extracting syllables from a word. Problems of syllabification are majorly caused from unknown and ambiguous words. This research aims to resolve these problems in Thai language by exploiting relationships among characters in the word. A character clustering scheme is proposed to generate units smaller than a syllable, called Thai Minimum Clusters (TMCs), from a word. TMCs are then merged into syllables using a trigram statistical model. Experimental evaluations are performed to assess the effectiveness of the proposed technique on a standard data set of 77,303 words. The results show that the technique yields 97.61% accuracy.

**Keywords:** Thai Syllabification, Thai Minimum Cluster, Trigram Model.

## 1 Introduction

Syllabification is a process to extract syllables from a word. This process is essential to many natural language processing tasks, especially for a text-to-speech system. Thai language with its unique characteristics both syntactically and semantically complicates the problem.

Many approaches were proposed to handle this task for Thai language, such as dictionary-based and rule-based methods. The idea is to group characters and produce syllables from the results. The main problem of Thai syllabification is how to handle unknown words, ambiguous words that can be differently pronounced, and proper names. For example, “กร” (korn), meaning a hand, can be pronounced as “กะ-ระ” (ka-ra) in a word “กรณี” (ka-ra-nee), meaning a case or a situation; or “รัตน” (rat-ta-na), meaning gems, can be pronounced as “รัต” (rat) in “สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี” (som-det-phra-tep-pha-rat-rat-su-da-sa-yam-bo-rom-ma-rat-cha-ku-ma-ree), the name of a Thai princess.

This paper proposes a novel technique to resolve these problems by using minimum character clustering and a trigram statistical model. The minimum character clustering technique is used to reduce time spent in cluster segmentation by grouping consecutive characters into clusters whose characters cannot be split further. These clusters will then be used in segmentation process instead of iterating though each

# Automatic Generation of a Pronunciation Dictionary with Rich Variation Coverage Using SMT Methods

Panagiota Karanasou and Lori Lamel

Spoken Language Processing Group, LIMSI-CNRS  
91403 Orsay, France  
{pkaran, lamel}@limsi.fr

**Abstract.** Constructing a pronunciation lexicon with variants in a fully automatic and language-independent way is a challenge, with many uses in human language technologies. Moreover, with the growing use of web data, there is a recurrent need to add words to existing pronunciation lexicons, and an automatic method can greatly simplify the effort required to generate pronunciations for these out-of-vocabulary words. In this paper, a machine translation approach is used to perform grapheme-to-phoneme (g2p) conversion, the task of finding the pronunciation of a word from its written form. Two alternative methods are proposed to derive pronunciation variants. In the first case, an n-best pronunciation list is extracted directly from the g2p converter. The second is a novel method based on a pivot approach, traditionally used for the paraphrase extraction task, and applied as a post-processing step to the g2p converter. The performance of these two methods is compared under different training conditions. The range of applications which require pronunciation lexicons is discussed and the generated pronunciations are further tested in some preliminary automatic speech recognition experiments.

**Keywords:** pronunciation lexicon, G2P conversion, SMT, pivot paraphrasing.

## 1 Introduction

Grapheme-to-phoneme conversion (g2p) is the task of finding the pronunciation of a word given its written form. Despite several decades of research, it remains a challenging task with many applications in human language technologies. Predicting pronunciations and variants, that is, alternative pronunciations observed for a linguistically identical word, is a complicated problem that depends on a number of diverse factors such as the linguistic origin of the speaker and of the word, the education and the socio-economic level of the speaker and the conversational context. Several approaches have been proposed in the literature to generate pronunciations. The simplest technique is manual creation, often relying on dictionary look-up in multiple resources, but making a pronunciation