Automatic Word Clustering in Studying Semantic Structure of Texts

Olga Mitrofanova

St. Petersburg State University, Faculty of Philology and Arts,
Department of Mathematical Linguistics,
Universitetskaya emb., 11
199034 St. Petersburg, Russia
{alkonost-om@yandex.ru}

Abstract. The purpose of the study is to prove that results of automatic word clustering (AWC) may contribute much in investigating semantic structure of texts and in evaluating plot complexity. Experiments were carried out for Russian texts, mainly stories and short novels. Data obtained in course of study allowed to formulate and verify several linguistic hypotheses.

Keywords: Automatic Word Clustering, Russian Corpora, Semantic Structure of Texts

1 Introduction

Formalization of text structure and quantitative evaluation of semantic relations between text units prove to be of considerable importance in various fields of natural language understanding: modelling plot structure, text summarization, evaluation of translation adequacy in parallel texts, automatic text indexing, classification of texts in corpora, etc. (for a detailed analysis cf. [1], [2]).

One of the procedures providing linguistic data on semantic structure of texts is automatic word clustering (AWC). It is assumed that AWC results help to reveal semantic structure of texts and to determine plot complexity. To prove this assumption, AWC procedure was carried out with the help of a specialized AWC toolkit based on word space model. Experimental procedure implied processing Russian texts, mainly stories and short novels. A set of key words describing major topics of the plot was assigned to each text, clusters of words with similar distributions were created for each key word. Data extracted from texts through AWC procedure admit thorough linguistic interpretation. Further comparison of cluster content and structure allowed to distinguish texts characterized by a plot including a dominating topic with a number of subtopics and texts characterized by a plot including a set of major (independent or correlating) topics.

Advances in Computational Linguistics.
Research in Computing Science 41, 2009, pp. 27-34

Received 07/12/08
Accepted 11/12/08
Final version 05/02/09

2 AWC Procedure

From a linguistic point of view, AWC is based on the possibility of detecting semantic similarity of words by comparing their syntagmatic properties (co-occurrence or distribution analysis); from a technical standpoint, AWC involves construction of vector-space models for processed texts; it means that the sets of contexts for each word are represented as distribution vectors in *N*-dimensional space [5], [7].

It is possible to evaluate semantic similarity of words by measuring distances between their distribution representations. Numerous metrics are used for the given purpose. The selection of metrics often depends on qualitative parameters of processed texts. In our case, cosine measure (*Cos*) was chosen as a basic metric. Results of measuring semantic distances are applied in clustering: words having similar distribution representations as a rule reveal similarity of meaning and should be included into the same cluster.

General approaches to clustering are exposed in hierarchical (agglomerative, divisive), partitioning (*K*-means, *K*-medoid, etc.), hybrid algorithms. Certain linguistic tasks require application of special clustering techniques, e.g. CBC [4], MajorClust [6], etc. The choice of a particular algorithm is determined by experimental conditions (corpora size, required speed of clustering, constraints for the size of resulting clusters, etc.). In our research preference was given to agglomerative clustering algorithm as it seems to be applicable in case of limited data and appropriate for processing texts of small / medium size.

Experiments were carried out with the help of AWC tool [3]. Python-based AWC software maintains procedures of text preprocessing and agglomerative clustering. Such parameters as names of input files (processed texts and key words describing the content of a text), context window size, weight assignment for context items, size of clusters, etc. are determined by users.

Text preprocessing is performed at the first stage. Context segmentation is carried out in accordance with a particular context window size. Automatic weight assignment may be done for lexical items taking into account their positions in contexts. Then, distribution representations of words are formed, co-occurrence matrix is built, semantic distances are calculated at the second step. These data are necessary for agglomerative clustering which is performed at the third step. An output file contains clusters of words with similar distributions in a text, such clusters being formed for each key word.

3 Linguistic Data

Experiments were carried out for over 20 Russian texts, mainly stories and short novels (cf. table 1). The texts differ in authorship (A. Belyaev, M. Bulgakov, N. Gogol, A. Grin, E. Zamyatin, A. Žitinsky, etc.), in size (N = 8 491 … 37 217 tokens), in lexical diversity (number of unique words L = 3 038 … 6 144 tokens, Somers coefficient $S = \ln \ln L / \ln \ln N$ = 0.920 … 0.936). In some experiments both raw and morphologically tagged texts were subjected to analysis. Processing raw texts provides data on distribution of word forms (tokens), while processing morpholo-

gically tagged texts allows to reveal interrelations between words (lemmas) within texts. In particular cases original Russian texts and their translations were considered as well. The texts were extracted from M. Moškov digital library (http://lib.ru/). Frequency lists for each text were created, additional statistical information (frequencies of words from various parts of speech, average sentence length, amount of dialogues, etc.) was obtained with the help of FantLab linguistic processor (http://www.fantlab.ru/).

Table 1. Texts subjected to analysis.

Author, title	**Size (tokens) (*N*)**	**Number of unique words (*L*)**	**Somers coefficient (*S*)**
Gogol N. *Taras Bul'ba*	37 217	6 144	0.920
Žitinsky A. *Časy s variantami* (*A Clock with Variants*)	28 092	5 197	0.922
Belyaev A. *Poslednij čelovek iz Atlantidy* (*The Last Man of Atlantis*)	26 892	5 160	0.924
Bulgakov M. *Sobačje serdce* (*Dog's Heart*)	25 218	5 321	0,928
Bulgakov M. *Rokovyje jajca* (*The Fatal Eggs*)	21 199	5 084	0.933
Zamyatin E. *Na kuličkah* (*In Kulički*)	20 832	4 544	0.928
Grin A. *Alyje parusa* (*Crimson Sails*)	20 366	4 984	0.933
Grin A. *Priključenija Ginča* (*Ginč's Adventures*)	19 120	5 017	0.936
Belyaev A. *Večny hleb* (*Eternal Bread*)	17 103	3 640	0.924
Grin A. *Kolonija Lanfier* (*Lanfier Colony*)	15 532	3 943	0.932
Belyaev A. *Mertvaja golova* (*A Dead Head*)	14 820	3 519	0.928
Gogol N. *Povest' o tom, kak possorilis' Ivan Ivanovič s Ivanom Nikiforovičem* (*A Tale of How Ivan Ivanovič Quarrelled with Ivan Nikiforovič*)	14 052	3 071	0.923
Belyaev A. *Zolotaja gora* (*A Golgen Hill*)	12 505	3 008	0.927
Belyaev A. *Čelovek, kotoryj ne spit* (*A Sleepless Man*)	11 943	3 104	0.931
Gogol N. *Viy*	11 800	2 824	0.926
Bulgakov M. *Zapisky na manžetah* (*Notes on the Cuff*)	10 056	3 038	0.937
Belyaev A. *Ni žizn', ni sm'ert'* (*Neither Life nor Death*)	9 681	2 653	0.931
Bulgakov M. *Morphij* (*Morphia*)	8 491	2 493	0.934

4 Experimental Results

In course of experiments a set of five key words – frequent words describing major topics of the plot – was assigned to each text, e.g.:

Zamyatin E. *Na kuličkah* (*In Kulički*):
key words {*kapitan* (*captain*), *Tihmen'*, *Marus'a*, *Andrej*, *Šmit*};

Žitinsky A. *Časy s variantami* (*A Clock with Variants*):
key words {*žizn'* (*life*), *vrem'a* (*time*), *časy* (*watch*), *ded* (*grandfather*), *ja* (*I*)}.

Clusters of lexical items with similar distributions were created for each key word. The following parameters of clustering were chosen in the experiments: similarity measure – *Cos*, context window size – ± 5, size of clusters – 10 items, no weight assignment. Previously it was found out that AWC performed with such parameters provides quite reliable data. Resulting clusters contain words or word forms associated with key words in a text and ordered according to *Cos* values. Distances between key words and their nearest neighbours in clusters (*D*) and difference between D_{max} and D_{min} in clusters (*Var*) were calculated for each text.

Table 2. Example (1): clusters of word forms extracted for key words in texts.

Text: Bulgakov M. *Morphij* (*Morphia*); key words *Polyakov*, *doktor* (*doctor*), *otdelenije* (*department*), *pis'mo* (*letter*), *Marja*; cluster elements are ordered in accordance with *Cos* values				
Polyakov	***doktor* (*doctor*)**	***otdelenije* (*department*)**	***pis'mo* (*letter*)**	***Marja***
pripiska (*postscript*) 0.328	*krasu* (*beauty*) 0.390	*terapevtičeskoje*	*nelepoje* (*absurd*) 0.428	*Vlasjevna* 0.731
krupnymi (*large*) 0.293	*zastrelils'a*	(*therapeutic*) 0.589	*isteričeskoje* (*hysterical*) 0.349	*prolepetala* (*prattled*) 0.326
bukvami (*letters*) 0.293	(*shot himself*) 0.390	*doktoru* (*doctor*) 0.490	153 0.349	*dviženije* (*movement*) 0.320
smerti (*death*) 0.289	*užas* (*horror*) 0.388	*hirurgičeskoje* (*surgery*) 0.431	*sarkoma* (*sarcoma*) 0.309	*šlepnula* (*slapped*) 0.320
umer (*died*) 0.285	*takoj* (*such*) 0.388	*Pavlu* (*Paul*) 0.423	*duše* (*soul*) 0.299	*bormotala* (*muttered*) 0.320
krasu (*beauty*) 0.254	*jehala* (*drove*) 0.379	*zaraznoje* (*infectious*) 0.409	*načalo* (*beginning*) 0.295	*brauning* (*Browning*) 0.287
pomutneli (*dimmed*) 0.219	*umer* (*died*) 0.333	*deckoje* (*infant*) 0.382	*roždalos'* (*was borning*) 0.284	*zadeta* (*touched*) 0.281
mimoletnuju (*fleeting*) 0.219	*drožala* (*trembled*) 0.323	*akušerskoje* (*obstetric*) 0.374	*ležalo* (*lay*) 0.259	*cepko* (*firmly*) 0.281
slyšno (*audible*) 0.215	*doroga* (*road*) 0.231	*mašina* (*car*) 0.340	*razdražat'* (*annoy*) 0.259	*boleznenno* (*painfully*) 0.281
	lampoj (*lamp*) 0.231	*bol'šoj* (*big*) 0.272		

Table 3. Example (2): clusters of word forms extracted for key words in texts.

Text: Belyaev A. *Čelovek, kotoryj ne spit* (*A Sleepless Man*); key word *preparat* (*medicine*), cluster elements are ordered in accordance with *Cos* values
***preparat* (*medicine*)**
himiki (*chemists*) 0.259
gotovyj (*ready*) 0.259
prodažu (*sale*) 0.236
uničtožavšij (*destroying*) 0.233
polučils'a (*came out*) 0.227
obnaružili (*discovered*) 0.227
polipeptidy (*polypeptides*) 0.195
vypuskalo (*produced*) 0.169
najdeny (*found*) 0.163

It seems that cluster elements often correspond to essential features of objects, persons or events denoted by key words and somehow emphasized in a text.

Relations between cluster elements can be characterized as syntagmatic and / or paradigmatic, e.g. synonymy & attributive relation: *terapevtičeskoje* (*therapeutic*), *hirurgičeskoje* (*surgery*), *zaraznoje* (*infectious*), *deckoje* (*infant*), *akušerskoje* (*obstetric*) – *otdelenije* (*department*); meronymy: *otdelenije* (*department*) – *doktor* (*doctor*); person – actions: *Marja* – *prolepetala* (*prattled*), *šlepnula* (*slapped*), *bormotala* (*muttered*), phraseological units and compounds: *Marja* – *Vlasjevna* (first name & second name), etc. (cf. table 2).

Those relations can be properly described in terms of semantic roles and lexical functions, e.g. action *obnaružili* (*discovered*) – agent *himiki* (*chemists*), result

preparat (*medicine*) – attribute *gotovyj* (*ready*), *uničtožavšij* (*destroying*); action *prodažu* (*sale*) – theme *preparat* (*medicine*), etc. (cf. table 3).

Thus, AWC allows to reveal and analyze not only standard but also occasional relations between lexical items which may be specific for a particular text or a set of texts of the same author or dealing with the same topic.

In some tests clustering was performed in two modes: with weight assignment and without weight assignment. In most cases clusters contain similar elements – word forms (tokens) in raw texts or words (lemmas) in tagged texts. At the same time those words or word forms within clusters may be ordered differently as regards their *Cos* values. So, clusters may be similar in content, but they may differ in structure (cf. table 4). It should be noted that in experiments with weight assignment *Cos* values for nearest neighbours of key words in clusters (*D*) seem to be lower than in experiments without weight assignment.

Table 4. Example: clusters obtained in experiments with / without weight assignment.

Text: Gogol N. *Viy*; key word *bursak* (*seminarist*), cluster elements are ordered in accordance with *Cos* values	
Clustering without weight assignment	**Clustering with weight assignment**
***bursak* (*seminarist*)**	***bursak* (*seminarist*)**
sodrognuls'a (*shuddered*) 0.479	*sodrognuls'a* (*shuddered*) 0.436
pozelenevšije (*green*) 0.442	*pozelenevšije* (*green*)0.405
otstupivši (*having stepped aside*) 0.420	*holod* (*cold*) 0.371
vperil (*stared*) 0.379	*izumlenija* (*amuzement*) 0.364
holod (*cold*) 0.359	*mertvyje* (*dead*) 0.338
čuvstvitel'no (*perceptibly*) 0.299	*čuvstvitel'no* (*perceptibly*) 0.305
izumlenija (*amuzement*) 0.295	*sv'atoj* (*saint*) 0.222
žizni (*life*) 0.259	*probežal* (*run*) 0.205
sv'atoj (*saint*) 0.200	*žizni* (*life*) 0.176

We also considered clustering results obtained in course of processing raw texts and morphologically tagged texts. Correspondence of word forms (tokens) and words (lemmas) in clusters created for raw and morphologically tagged texts (cf. table 5) proves the existence of stable intrinsic relations underlying text structure. These relations remain almost intact as the analysis moves from the level of word forms (tokens) to the level of words (lemmas). So, AWC procedure may furnish us with additional information on the integrity and continuity of the text as a complex of heterogeneous linguistic units.

AWC proves to be of much use in comparative analysis of original texts and translations, as it often allows to evaluate stylistic and semantic similarity of texts. Similarity of clusters formed for a word and its translation equivalent reveals correspondence between contexts of those words in the original and in translation, while differences of content and structure of such clusters imply syntactic / morphological / lexical differences of texts in question as well as inconsistency in the choice of translation equivalents for a particular word or for lexical items co-occurring with this word in contexts (cf. table 6).

As statistical parameters of texts may influence results of clustering, additional tests were required. We've studied the texts written by A. Belyaev which reveal

common semantic structure and are characterized by a branching plot with numerous and frequently changing topics. The given texts differ in size and in number of unique words. At the same time, distances between key words and their nearest neighbours in clusters don't vary much for those texts ($D \in [0.088 \ldots 0.259]$). It turns out that such parameters as size and number of unique words play important but not decisive role in studying text structure by means of AWC.

Table 5. Example: clusters obtained in experiments with raw and tagged texts.

Text: Bestužev-Marlinsky A. *Strašnoje gadanje* (*A Scary Fortune-telling*); key word *neznakomec* (*stranger*), cluster elements are ordered in accordance with *Cos* values	
Clustering in a raw text (tokens)	**Clustering in a tagged text (lemmas)**
***neznakomec* (*stranger*)**	***neznakomec* (*stranger*)**
stenky (*wall*) 0.219	*drognut'* (*quaver*) 0.223
podjezdu (*entrance*) 0.219	***trost'* (*cane*) 0.198**
***vysadiv* (*having put off*) 0.219**	***vysadit'* (*pull off*) 0.197**
***večor* (*evening*) 0.218**	*zajti* (*overstep*) 0.196
***zahvatyvaja* (*seizing*) 0.216**	***večor* (*evening*) 0.196**
***rasseržen* (*angry*) 0.216**	*kalitka* (*gate*) 0.195
***trost'* (*cane*) 0.193**	***zahvatyvat'* (*seize*) 0.194**
***gorst'ami* (*in handfuls*) 0.188**	***rasserdit'*(*anger*) 0.192**
car'a (*tzar*) 0.172	***ironičeskij* (*ironical*) 0.173**
***ironičeskoju* (*ironical*) 0.165**	***gorst'*(*in handfuls*) 0.169**

Table 6. Example: comparison of clusters formed for test words in the original text and in translation.

Texts: Grin A. *Alyje parusa* (*Crimson Sails*); test words *Sekret* (*Secret*), *galiot* (*galliot*), cluster elements are ordered in accordance with *Cos* values			
Russian text	**English text**	**Russian text**	**English text**
Sekret (Secret)	***Secret***	***Galiot (Galliot)***	***galliot***
potr'asenija (*shock*) 0.239	*intimations* 0.213	*trehmačtovyj* (*three-mastered*) 0.800	*masted* 0.843
vdohnovennogo (*inspired*) 0.210	*hurries* 0.212	*dvesti* (*two hundred*) 0.700	*purchased* 0.556
dvesti (*two hundred*) 0.178	*agitation* 0.202	*šest'des'at* (*sixty*) 0.600	*sixty* 0.527
neuderžimymi (*uncontrollable*) 0.178	*rounding* 0.201	*kuplennyj* (*purchased*) 0.600	*ton* 0.509
slezami (*tears*) 0.178	*shock* 0.173	*tonn* (*ton*) 0.500	*brig* 0.316
nravits'a (*likes*) 0.149	*cape* 0.173	*Grejem* (*Gray*) 0.329	*hundred* 0.271
kamenistoj (*rocky*) 0.149	*uncontrollable* 0.144	*sobstvennikom* (*proprietor*) 0.3	*orion* 0.222
padajuš'im (*falling*) 0.147	*masted* 0.117	*kapitanom* (*captain*)0.291	*rugged* 0.211
golovokružitel'no (*astoundingly*) 0.117	*galliot* 0.093	*mačty* (*masts*) 0.290	*Arthur* 0.189

Thorough treatment of AWC results allowed us to distinguish three main types of texts with regard to their semantic structure (Types 1, 2, and 3).

Type 1 is represented by texts characterized by a plot including a dominating topic with a number of subtopics. For such texts distances between key words and their nearest neighbours in clusters (D) and difference between $D_{\max}$ and $D_{\min}$ in clusters (Var) are as follows: $D \geq 0.300$, $Var \geq 0.200$.

Type 2 is represented by texts characterized by a plot including a set of major (probably independent) topics. For such texts distances between key words and their nearest neighbours in clusters (D) and difference between $D_{\max}$ and $D_{\min}$ in clusters (Var) are as follows: $D < 0.300$, $Var < 0.200$.

Type 3 is represented by texts characterized by a plot including a set of major (probably correlating) topics. For such texts distances between key words and their nearest neighbours in clusters (D) and difference between D_{max} and D_{min} in clusters (Var) are as follows: $D \geq 0.300$, $Var < 0.200$.

Examples of texts representing Types 1, 2, and 3 are given in table 7.

Table 7. Texts representing Types 1, 2 and 3.

Type, author, title	***D***	***Var***
Type 1		
Gogol N. *Taras Bul'ba*	0.379	0.252
Grin A. *Priključenija Ginča* (*Ginč's Adventures*)	0.406	0.231
Gogol N. *Povest' o tom, kak possorilis' Ivan Ivanovič s Ivanov Nikiforovičem* (*A Tale of How Ivan Ivanovič Quarrelled with Ivan Nikiforovič*)	0.453	0.357
Gogol N. *Viy*	0.547	0.424
Belyaev A. *Zolotaja gora* (*A Golgen Hill*)	0.566	0.471
Bulgakov M. *Morphij* (*Morphia*)	0.731	0.403
Type 2		
Žitinsky A. *Časy s variantami* (*A Clock with Variants*)	0.149	0.048
Zamyatin E. *Na kuličkah* (*In Kulički*)	0.174	0.068
Grin A. *Alyje parusa* (*Crimson Sails*)	0.204	0.091
Bulgakov M. *Sobačje serdce* (*Dog's Heart*)	0.212	0.103
Belyaev A. *Ni žizn', ni sm'ert'* (*Neither Life nor Death*)	0.222	0.070
Belyaev A. *Poslednij čelovek iz Atlantidy* (*The Last Man of Atlantis*)	0.224	0.115
Grin A. *Kolonija Lanfier* (*Lanfier Colony*)	0.224	0.139
Belyaev A. *Mertvaja golova* (*A Dead Head*)	0.243	0.155
Belyaev A. *Čelovek, kotoryj ne spit* (*A Sleepless Man*)	0.259	0.144
Belyaev A. *Eternal bread* (*Večny Hleb*)	0.268	0.130
Bulgakov M. *Rokovyje jajca* (*The Fatal Eggs*)	0.279	0.182
Type 3		
Bulgakov M. *Zapisky na manžetah* (*Notes on the Cuff*)	0.359	0.091

Our observations on semantic structure of texts require more detailed consideration and further verification.

5 Conclusion

In course of our experiments performed for Russian stories and short novels we proved that AWC may be of great help in distinguishing three types of texts as regards their semantic structure. We managed to describe texts of different plot complexity: texts revealing a dominating topic and a set of subtopics, texts revealing a set of major (probably independent) topics, and texts revealing a set of major (probably correlating) topics. Linguistic analysis of cluster content and structure

allowed to study standard as well as occasional semantic relations between lexical items occurring in texts. Experiments on AWC performed for raw and morphologically tagged texts proved the existence of intrinsic relations underlying text structure, those relations being preserved at two levels of analysis: the level of word forms (tokens) and the level of words (lemmas). Comparison of AWC results obtained for the original texts and their translations proved to be relevant in the evaluation of stylistic and semantic similarity of texts.

Further research implies experiments carried out for texts of different size, genre and authorship, with expanded sets of key words, with changing parameters (context window size, cluster size, etc.).

Acknowledgements. The author expresses deep gratitude to CICLing–2009 Program and Organizing Committees for having accepted the paper, to reviewers for their kind attention and wise comments which helped to improve the paper, to Mikhail Alexandrov (Barcelona, Spain) and Elena Iagounova (St. Petersburg, Russia) for their deep interest in the given research and for their help and encouragement.

References

1. Bolshakov, I.A., Gelbukh, A.: Computational Linguistics: Models, Resources, Applications. IPN – UNAM – Fondo de Cultura Económica (2004)
2. Leontjeva, N.N.: Avtomatičeskoje Ponimanije Tekstov: Sistemy, Modeli, Resursy. Moscow (2006)
3. Mitrofanova, O., Mukhin, A., Panicheva, P., Savitsky, V.: Automatic Word Clustering in Russian Texts. In: Matoušek, V., Mautner, P. et al. (eds.): Text, Speech and Dialogue. Proceedings of the Tenth International Conference TSD–2007, Pilsen, Czech Republic, September 3–7, 2007. Lecture Notes in Artificial Intelligence, Vol. 4629. Springer-Verlag, Berlin Heidelberg New York (2007) 85–91
4. Pantel, P.: Clustering by Committee. Ph.D. Dissertation, Department of Computing Science, University of Alberta (2003) http://www.isi.edu/~pantel/Content/publications.htm
5. Sahlgren, M.: The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-dimensional Vector Spaces. Ph.D. dissertation, Department of Linguistics, Stockholm University (2006) http://www.sics.se/~mange/TheWordSpaceModel.pdf
6. Stein, B., Meyer zu Eissen, S.: Document Categorization with MajorClust. In: Proceedings of the 12th Workshop on Information Technology and Systems WITS–02. Barcelona, Spain (2002) 91–96
7. Widdows, D.: Geometry and Meaning. Center for the Study of Language and Information – Lecture Notes, Vol. 172. The University of Chicago Press (2004)