

What's your style?

Automatic genre identification with neural network

Andrea Dömötör^{1,3}, Tibor Kákonyi¹, and Zijian Győző Yang^{1,2}

¹ MTA-PPKE Hungarian Language Technology Research Group

² Pázmány Péter Catholic University, Faculty of Information Technology and Bionics

³ Pázmány Péter Catholic University, Faculty of Humanities and Social Studies
{yang.zijian.gyozo, domotor.andrea}@itk.ppke.hu
kakonyi.tibor@hallgato.ppke.hu

Abstract. Genre identification is an important task in natural language processing which can be useful for many practical and research purposes. However this task is extremely hard because genre is not a homogeneous and unequivocal property of texts and it is sometimes barely separable from the topic. In this paper we compare the performance of two different automatic genre identification methods. We classified six text types: literary, academic, legal, press, spoken and personal. In one part of our research we did experiments with traditional machine learning methods using linguistic, n-gram and error features. In the other part we tested the same task with a word embedding based neural network. In this part we experimentalised with different training data (words only, POS-tags only, words and POS-tags etc.). Our results revealed that neural network is a suitable method for this task while traditional machine learning showed significantly lower performance. We gained high (around 70%) accuracy with our word embedding based method. The results of the different text categories also showed differences which is related to the stylistic properties of the studied genres.

Keywords: genre identification, text classification, machine learning, neural networks, word embedding, stylistics

1 Introduction

Automatic genre identification is an application of computational stylistics which originates from the idea that the different text types have different lexical and grammatical features.

However, the term *genre* can be interpreted in several ways (see overview in [2]), modern definitions usually mention the communicative purpose (function), content and form as the main distinctive properties of genres. In this study we concentrate on the last characteristic: the form, what is to say, the structural and lexical features of the different text types. This decision is in line with both our methods and motivation. Firstly, we used linguistic properties

(words, lemmata and POS-tags) as training data in our experiments. On the other hand, our purpose of building an automatic genre identification system is also linguistically motivated. We expect this system to support the creation of genre-specific (sub)corpora which can be useful for corpus linguistic and stylistic studies. Genre identification may also help other natural language processing systems (for example, rule-based parsers) by allowing the use of genre-specific rules.

The traditional genre identification methods are based on the selection of features ([5][9]). These can be either surface features like function words, genre-specific words, word length or sentence complexity; structural features, for example parts of speech or verb tenses or presentation and other features, such as token type or links. The classification algorithms used for this task also vary in the literature from decision trees, through naive Bayes and regression to neural networks and clustering.

In this paper we compare two methods on the same training data set. In one part of this study, we experimentalized a classification model based on feature extraction. In the other part we used deep neural networks and word embedding. The peculiarity of our work is that it is sentence-based, while other studies of genre identification usually use bigger text units. (However, not all of them. [6] for instance, actually searches the minimal unit necessary to identify genre.) The reason we chose this type of task is, on one hand, that the style of web pages may not be homogeneous. For this reason it is important to be able to deal with smaller text units in order to build genre-specific corpora. On the other hand, as we mentioned before, the study also has the purpose to enable genre identification for natural language processing tools and corpus linguists. In these cases it can be necessary to identify the genre of a one-sentence input or research data.

2 Training data

The training data was extracted from the Hungarian Gigaword Corpus (HGC) [8]. The corpus contains 187.6 million tokens of lemmatized and morphologically annotated texts from different genres. The analysis of the corpus was realized with the Humor morphological analyzer tool [10] which is a reversible, string-based, unification approach for lemmatizing and disambiguation.

Our training data was provided by the press, literary, academic, legal, personal and spoken language subcorpora. The press subcorpus contains texts from news webpages. This adds up the major part of the whole HGC. The literary subcorpus is a processed collection of digitally available texts of Hungarian literature. The academic texts originate from a Hungarian digital library. The legal subcorpus contains texts of laws, decrees and parliamentary records. The personal subcorpus is built of web forum conversations. These texts are usually below standard and often noisy. Finally, the spoken language corpus consists of transcriptions of radio programmes.

We queried 300 thousand random sentences from each type. The training data elaborated of these sentences contains the original words, lemmata and POS-tags. We used all these three characteristics for our experiments because we presume that genres have both particular lexical and structural characteristics.

Vocabulary is an obvious distinctive feature of text types. Table 1 shows the most frequent trigrams of the different genres (not taken into consideration punctuation marks and conjunctions). As it can be seen, the categories are more or less recognizable from their common collocations, however there are similarities due to similar topics (legal, press, spoken) or to the generality of the genre’s vocabulary (personal, literary). As Hungarian is a morphologically rich language, it seems adequate to use both full word forms and lemmata.

The relevance of POS-tags is demonstrated in Table 2 which shows the relative frequency of personal pronouns in each text type. These data reveal that press and academic texts show strong preference to the third person⁴, while the second person is slightly more prominent in personal and literary texts compared to the other genres. These characteristics are expected to cause significant differences in the distribution of (verbal) POS-tags.

We created 5 different kinds of training and test corpus. These contain the following information:

- Full word forms (original text)
- Lemmata
- POS-tags
- Full word forms and lemmata
- Full word forms and POS-tags

The combined types are necessary to distinguish homographs. The two missing types (lemmata and POS-tags; full word forms, lemmata and POS-tags) are redundant, because the combinations of full word forms and POS-tags, lemmata and POS-tags and full word forms, lemmata and POS-tags equally determine the word unambiguously.

We used the texts as they appeared in the corpus, no preprocessing steps or normalization was applied. In our judgment quality issues can play a significant role in genre identification, for instance, the omission of accented characters or punctuation marks is a characteristic of informal texts. The only intervention to the corpus data was the filtering of duplications and "trash" (like html tags or meta data).

3 Methods and experiments

3.1 Traditional machine learning method

In one part of our research we did experiments to build a classification model using traditional machine learning. For this task we tested various classification

⁴ This stands for legal texts as well, if we take into consideration that the formal *you* (*ön*) in Hungarian also takes the third person.

Table 1. Most frequent trigrams of text types

Personal
nem csak a – 'not only the'
az a baj – 'the problem is'
a mai napon – 'this day'
még akkor is – 'even if'
még mindig nem – 'still not'
Legal
megadom a szót – 'I give the floor'
az Európai Unió – 'the European Union'
a módosító javaslatot – 'the amendment'
köszönöm a szót – 'thank you for the floor'
nem fogadta el – 'has not accepted'
Literary
ez volt a – 'this was the'
ha nem is – 'even if not'
még akkor is – 'even if'
még mindig nem – 'still not'
nem is tudom – 'I don't know'
Spoken
én azt gondolom – 'I think'
az Európai Unió – 'the European Union'
jó reggelt kívánok – 'good morning'
jó napot kívánok – 'good afternoon'
az Európai Bizottság – 'the European Committee'
Press
az Egyesült Államok – the United States
az Európai Unió – the European Union
a tervek szerint – 'according to plans'
a múlt héten – 'last week'
az Európai Bizottság – 'the European Committee'
Academic
a második világháború – 'the second world war'
a 19. század – 'the 19th century'
részt vett a – 'took part in'
volt az első – 'was the first'
a 20. század – 'the 20th century'

methods and the Random Forest algorithm gained the best results, thus in this paper we show only the results of our Random Forest model (RFM).

To build the RFM, we used the PiRate system [12]. We implemented 37 different kinds of features. According to the functionality, we can separate these features into the following categories:

- linguistic features:

Table 2. Relative frequency of pronouns in different genres

	én ('I')	te (you.sg) (informal)	ön (you.sg) (formal)	ő ('he/she')	mi ('we')	ti (you.pl)	ők ('they')
Personal	33.3%	15.1%	3.0%	23.2%	11.2%	3.9%	10.4%
Legal	28.5%	0.6%	26.0%	19.7%	16.2%	0.2%	8.7%
Literary	32.9%	10.7%	1.2%	32.1%	10.6%	1.5%	11.0%
Spoken	30.2%	1.6%	9.1%	26.0%	18.3%	0.4%	14.5%
Press	14.1%	2.8%	3.2%	41.9%	16.3%	0.5%	22.3%
Academic	11.1%	3.5%	0.9%	53.3%	8.4%	0.9%	21.8%

- percentage of nouns, verbs, pronouns, adverbs, adjectives, conjunctions, pronouns, determiners, preverbs, numerals, interjections in the sentence;
- ratio of number of nouns and verbs in the sentence;
- ratio of number of nouns and adjectives in the sentence;
- ratio of number of verbs and preverbs in the sentence;
- ratio of number of nouns and determiners in the sentence;
- number of tokens;
- average word length in the sentence;
- n-gram features:
 - sentence LM probability;
 - sentence LM perplexity;
 - LM probability of lemmas and POS tags of the sentence;
 - LM perplexity of lemmas and POS tags of the sentence;
- neural network features:
 - 1-gram, 2-gram and 3-gram perplexity;
- error features:
 - percentage of accented words in the sentence;
 - percentage of unknown words in the sentence;
 - percentage of punctuation marks in the sentence.

The training of the n-gram models (for the n-gram features) was effectuated with the SRILM [11] toolkit. As n-gram training corpus we used a subcorpus of the HGC that contains 98500 lemmatized and POS-tagged sentences.

For the training of the neural network language model we used a subcorpus of the Pázmány Corpus [3] that contains 1 million sentences. The language model was built with an RNN architecture with GRUs (Gated recurrent unit). We also used a Hungarian word embedding model [7] for word representation.

3.2 Word embedding based neural network method

In the other part of our research we made experiments using fastText, which is a state-of-the-art, neural network based library for word embedding [4] and text classification [1] developed by Facebook Artificial Intelligence Research.

For text classification fastText uses a linear classifier based on supervised learning, it needs labeled corpora as training and validation sets. During the

training fastText builds an embedding model where labeled sentences and labels are represented as vectors in a way that a sentence is really close to its associated labels in the vector space.

An initial sentence vector is the average of embedding vectors of words inside the sentence. (An advantageous ability of fastText is that it does not work simply with words but with n-gram features, hence it is able to handle some partial information about the local word order.) The sentence vector is fed into a linear classifier and softmax function is used to calculate the probability distribution over labels. fastText uses stochastic gradient descent algorithm to maximize the probability of the correct label belonging to the sentence.

In our experiment each sentence of the corpus had one label that marked which style that piece of text belongs to. We trained models for all five kinds of the corpus with 27 different parameter sets that are generated as combinations of the following values:

- number of epochs (number of times fastText sees a training example): 5, 27 or 50;
- learning rate (degree of the model’s change after processing an example): 0.1, 0.5 or 1.0;
- maximal length of word n-grams: 1, 3 or 5.

(Only the model giving the best results is mentioned for each corpus variety in the Results section.)

4 Results and Evaluation

Table 3 shows the accuracy results of our experiments. First, comparing the performance of the different training corpus types in the method described in chapter 3.2 it can be seen that, as was expected, the only POS-tag version gave the lowest results, 52% in average and 56.6% best case (0.1 learning rate, 5 epochs, 5-grams). Nevertheless, these results are still remarkable, taking into consideration how limited information the model had, and even though it performed far above random. This means that the studied genres do have unique structural properties and the difference between them is not only lexical or thematic. As for the other four types of subcorpora, we have almost the same results and they also share the best parameter set (0.1 learning rate, 5 epochs, 3-grams). It seems, contrarily to the assumptions, that full morphology does not contribute much to the lexical-based genre identification: the model works just the same with only lemmata as with full word forms and POS-tags. In all four cases we got a fair (around 70% best case) result. Other observation worth to mention is that the increase of number of epochs did not prove to increase the performance.

Table 3 also shows that our Random Forest Model performed significantly below fastText, even if the last one only had the POS-tags as input data. These results demonstrate that word embedding methods are much powerful for this task than traditional machine learning. The relative inefficiency of the Random

Table 3. Accuracy results of the word embedding based and the random forest method

	Average accuracy	Best accuracy	Best n-gram parameter
Words	68.5%	70.7%	3
Lemmata	68.3%	70.7%	3
Words + POS-tags	68.2%	70.3%	3
Words + lemmata	68.0%	70.1%	3
POS-tags	52.0%	56.6%	5
RFM	-	43.2%	-

Forest Model is, however, not that surprising if we consider that the majority of the features used in this method is not sensitive to the vocabulary.

We also measured precision, recall and F-score by category (Table 4). In this case the four subcorpus types that contained words or lemmata still performed almost the same in the word embedding based method, for this reason we only show the results of the models using words and POS-tags.

As seen, fastText’s full word form measurement gave the best result for legal texts with an F-score over 80%. Apparently, this is the genre with the most characteristic vocabulary, which is presumably related to its thematical boundedness. Literary, academic and spoken texts also achieved high points with this method. The relatively low performance of the personal type can be attributed to the low quality of this subcorpus. This assumption is even more plausible considering the fT-POS results. The difficulty of identifying this kind of texts by POS-tags can be caused by the significant number of erroneous tags (which occur frequently in this subcorpus due to the omission of accents, typos, abbreviations etc.).

The fT-POS results follow the same order but the numbers are lower in proportion (except for the extremely low recall of the personal type).

The majority of Random Forest Model’s results does not even reach the f-measure of word embedding with POS-tags, except in case of personal texts. The scores gained by the traditional machine learning method are generally low. The highest f-measure (50.8%) belongs to the literary genre but this result is still lower than the worst score of fT-word.

To detect the common faults we made a confusion matrix of the fastText-word experiment (Table 5). The personal type is often confused with the literary. The reason perhaps is that both genres are quite liberal in terms of text composition. The spoken texts seem to be related with the press genre. This can be explained with the similarity of their topics. As we mentioned before, the spoken subcorpus consists of transcriptions of radio programmes which often contain news and public topics. The relatively high number of confusions between the press and academic genres may be explicable with the observation shown in Table 2 that these text types typically prefer the third person.

Finally, it should be mentioned that the task of genre identification by definition does not assume 100% accuracy, as genre is not a unequivocal property of texts. Any genre can contain neutral sentences which have no distinctive stylis-

Table 4. Precision and recall results by category

		fT-word	fT-POS	RFM
Legal	Precision	82.47%	62.01%	45.9%
	Recall	79.52%	65.82%	40.2%
	F-Measure	80.96%	63.86%	42.9%
Literary	Precision	72.08%	57.35%	44.9%
	Recall	79.00%	67.01 %	58.4%
	F-Measure	75.38%	61.80%	50.8%
Academic	Precision	74.38%	59.44%	45.2%
	Recall	71.75%	61.55%	53.6%
	F-Measure	73.04%	60.48%	49%
Spoken	Precision	69.97%	54.57%	36.1%
	Recall	72.23%	55.81%	37%
	F-Measure	71.08%	55.19%	36.5%
Press	Precision	56.47%	43.46%	36.8%
	Recall	57.58%	41.53%	33.4%
	F-Measure	57.02%	42.48%	35%
Personal	Precision	57.72%	53.16%	52.6%
	Recall	48.00%	25.17%	37%
	F-Measure	52.41%	34.16%	43.3%

Table 5. Confusion matrix

	Personal	Legal	Literary	Spoken	Press	Academic
Personal	57.72%	3.88%	15.41%	7.10%	9.40%	6.50%
Legal	1.94%	82.47%	1.94%	4.95%	4.92%	3.79%
Literary	7.92%	2.30%	72.08%	4.87%	5.92%	6.90%
Spoken	4.55%	6.73%	4.95%	69.97%	10.27%	3.52%
Press	8.79%	6.25%	5.18%	12.79%	56.47%	10.53%
Academic	3.87%	3.64%	6.59%	2.92%	8.59%	74.38%

tic characteristics. Therefore, 70% accuracy on sentence level can be considered significant.

5 Conclusion

In this paper we compared the results of a traditional machine learning and a word embedding based method in the task of automatic genre identification. For both methods we used corpora that contained lexical and grammatical information, namely words, lemmata and POS-tags. According to our results, the word embedding method is much more powerful for this task. The performance of the neural network based system far surpassed the traditional machine learning algorithms. With word embedding we achieved promising results (around 70% accuracy).

Our experiments provided other interesting findings as well. The word embedding measurements revealed that using the POS-tags only can be more effective

than expected. This suggests that genres have specific structural characteristics which allow to identify them without lexical or topic-related features.

Other observation of linguistic interest is that we got the same result when using full word forms and lemmata despite that Hungarian is an agglutinative language which means that a lemma can have varied word forms.

As for genre-related results, we found that legal, literary and academic texts are easier to identify than the other three examined genres (spoken, press, personal). It seems that these text types have more representative lexical and structural characteristics than the others. It is also important to remark that the spoken and personal language types represent greater variation in topics which makes the lexical-based genre identification harder.

Finally, it is to be mentioned that the traditional machine learning methods are more language-dependent than word embedding. The feature set of our machine learning model contains features that are specific for Hungarian (like the number of accented characters). Other languages may need different features. However, the word embedding method can be used to any language without modifications.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. CoRR (2016)
2. Clark, M., Ruthven, I., O'Brian Holt, P.: The evolution of genre in wikipedia. *Journal for Language Technology and Computational Linguistics* 24(1), 1–22 (2009)
3. Endrédi, I., Prószték, G.: A pázmány korpusz. *Nyelvtudományi Közlemények* (112), 191–206 (2016)
4. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. CoRR (2016)
5. Lustrek, M.: Overview of Automatic Genre Identification. Joef Stefan Institute, Department of Intelligent Systems, Ljubljana, Slovenia (2007)
6. McCarthy, P.M., Myers, J.C., Briner, S.W., Graesser, A.C., McNamara, D.S.: A psychological and computational study of sub-sentential genre recognition. *Journal for Language Technology and Computational Linguistics* 24(1), 23–56 (2009)
7. Novák, A., Novák, B.: Magyar szóbeágyazási modellek kézi kiértékelése. In: XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018). Szegedi Tudományegyetem, Szeged, Hungary (2018)
8. Oravecz, C., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Calzolari, N., et al. (eds.) *Proceedings of the 9th International Conference on Language Resources and Evaluation. ELRA*, Reykjavik, Iceland (may 2014)
9. Petrenz, P., Webber, B.L.: Stable classification of text genres. *Computational Linguistics* 37(2), 385–393 (2011)
10. Prószték, G., Tihanyi, L.: Humor – a morphological system for corpus analysis. In: *Proceedings of the first TELRI seminar in Tihany*. pp. 149–158. Budapest, Hungary (1996)
11. Stolcke, A.: Srilm – an extensible language modeling toolkit. In: *IN PROCEEDINGS OF THE 7TH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (ICSLP 2002*. pp. 901–904 (2002)

12. Yang, Z.G., Laki, L.J.: rate: A task-oriented monolingual quality estimation system. *International Journal of Computational Linguistics and Applications* 8 (2017)