An Online System for Neologism Detection in Bulgarian

Ivelina Stoyanova, Svetlozara Leseva, Martin Yalamov, Svetla Koeva
Department of Computational Linguistics,
Institute for Bulgarian Language, Bulgarian Academy of Sciences
E-mail: {iva,zarka,martin,svetla}@dcl.bas.bg

Abstract

We propose a combined method for lexical neologism detection. The method employs several resources and techniques for identification, filtering and ranking of neologism candidates: linguistic annotation; exclusion lists compiled from lexicographic resources and text corpora; named entity recognition and spelling error detection; grouping of candidates using stemming to cater for morphological variations. The set of potential neologisms are subsequently ranked based on their frequency and a dispersion measure which accounts for the candidates' distribution with respect to: (a) documents; (b) sources; (c) domains; and (d) period of time. The method is integrated in an online chain for collection and processing of texts in Bulgarian which is used for media monitoring and lexical and lexicographic analysis.
Keywords: online lexical monitoring; neologism detection; neologism ranking; Bulgarian

1. Introduction

Recognition and description of the properties of lexical changes is still a challenge for modern lexicography and Natural Language Processing (NLP).

Neologisms are usually divided into two categories: lexical and semantic. Lexical neologisms are new graphical words that refer to novel or known concepts. Semantic neologisms are newly developed senses of existing words. The automatic identification of lexical and semantic neologisms differs to the extent that the former investigates changes related to the neologism candidates, while the later studies changes in their collocational environment (Renouf, 1993).

In this paper we propose a combined method for neologism detection (including both single-words and multiword expressions) which employs various resources and techniques. The method is integrated in a web-based system for monitoring and analysis of Bulgarian media content and aims at facilitating lexicographic work in the identification of new words.

In the next section we briefly present the most common approaches employed in similar tasks. Section 3 describes the method adopted in our research. Section 4 outlines the workflow and user interface of the system for automatic neologism detection from media content. We conclude by sketching the envisaged future work.

2. Methods for Detection of Vocabulary Changes

The main methods for neologism detection are based on: (i) application of language resources, such as exclusion lists of 'known' words or pattern matching based on lexical cues; (ii) statistical measures or machine learning applied to corpora

containing samples from a relatively large period (diachronic methods); (iii) a combination of the two.

The methods based on exclusion lists, such as the one proposed by O'Donovan (2008), use word lists compiled from existing lexicographic resources, such as dictionaries or corpora, combined with filters for the elimination of non-words, typographical errors, named entities (NEs). The pattern-based methods (Paryzek, 2008) rely on the so-called lexical cues – markers of lexical novelty and/or punctuation marks that usually signal the proximity of new words.

Stenetorp (2010) proposes an SVM based machine-learning method for extraction of Swedish neologisms. The features include the number of occurrences, the points in time marking the first and the last occurrence, lexical cues, presence in a dictionary, and the ability of a spellchecker to found correction suggestions.

A combined method is proposed by Falk et al. (Falk et al., 2014). The unrecognised words in a corpus are filtered using dictionary derived word lists and a list of named entities. A classifier is trained to recognise new words on the basis of manually validated neologisms in the set of unrecognised words in the corpus.

Kilgarriff (2015) present a system for neologism detection, Diacran, which is implemented within the Sketch Engine. The system uses subcorpora reflecting prominent 'time slices'. For each word whose graph line has a high gradient and high correlation and whose overall frequency is high, the authors estimate the combined score of these three factors to obtain an overall score for the word (the highest combined scores point to the 'most interesting' words).

Kerremans et al. (2012) performs direct automatic analysis of web pages retrieved through web crawling. Their content is cleaned, processed and filtered and a list of possible neologisms is presented to linguists for analysis. The method is implemented in a system called NeoCrawler.

The analysis shows that although comprehensive methods have been developed for identification of new words, they are rarely implemented into fully functional system for online observation and comparison. To the best of our knowledge, no system for new words detection exists for Bulgarian.

3. A Combined Method for Neologism Detection

We propose a combined method for neologism detection which employs several techniques for the purpose of detecting and ranking neologism candidates:

(1) linguistic annotation and initial identification of candidates;

(2) application of exclusion lists compiled from (i) lexicographic resources; and (ii) text corpora;

(3) filtering of improbable candidates using named entity recognition and spelling error detection;

(4) grouping of candidates using stemming to cater for morphological variations;

(5) filtering and ranking of candidates using statistical analysis with a view to refining the output and facilitating manual verification.

3.1. The Corpora

We use two sets of data from the Bulgarian National Corpus (BulNC) (Koeva et al., 2012). The BulNC is chosen because it contains a large variety of texts of different size, media type (written and spoken), style, period, and languages and totals 5.4 billion tokens, of which 1.2 billion constitute the Bulgarian part of the corpus. The BulNC combines the properties of static corpora (a detailed metadata classification), dynamic (monitor) corpora (continuous expansion) and opportunistic corpora (collection of as much linguistic data as possible). The corpus is expanded mainly automatically with texts crawled from the Internet.

The first data set is a corpus of texts created during an earlier reference period (currently, from 1945 until the end of 2014). The second one comprises the most recent texts harvested from various media sources. We allow over a year between the reference period and the observed period in order to be able to observe the adoption of new words into the language as a process rather than a single event.

3.2. Linguistic Annotation and Initial Identification of Candidates

The data set studied for potential neologisms is processed and annotated using the Bulgarian Language Processing Chain (Koeva & Genov, 2011). The wordforms whose part of speech cannot be guessed by the tagger (and are therefore tagged M – miscellaneous) and those that are not found in the inflectional dictionary and are therefore assigned a basic POS tag (such as N, V, A), constitute the initial set of unrecognised words. Those that contain characters other than Cyrillic letters and a dash are additionally filtered out.

The method is also adapted to identify possible multiword (MWE) neologisms and follows the same steps as for single-words. Currently, we only consider MWEs of two components. In order to extract bigrams we apply: (a) a syntactic filter to identify candidates of certain types of structure; and (b) statistical analysis to select candidates with association measure (MI – mutual information) above a threshold. Further, we follow the same procedures as for the single-words.

3.3. Exclusion Lists

At the next step, extensive exclusion lists of known words are applied to the lists of candidates obtained at the preprocessing stage. The exclusion lists are compiled

from two sources. On one hand, these are static lexicographic resources – various dictionaries, indices from books, lists of words with spelling errors, lists of named entities, etc. A total of 41 exclusion lists are applied, including more than 600,000 unique wordforms. On the other hand, we compile an exclusion list from the BulNC containing additional 1.5 million unique wordforms. The lists are continuously updated with candidates that have been manually rejected as new words, as well as with words that over time became established in the language and are no more considered as new words. Exclusion lists for MWEs comprise dictionaries of MWEs, as well as all bigrams in the BulNC.

3.4. NE Recognition and Spelling Error Identification

The first most frequent candidates after the application of the exclusion lists are usually words with wrong spelling and unfamiliar names (not in the dictionaries and exclusion lists). We apply a simple NE recognition technique where we identify words appearing with initial capital letter in mid-sentence as likely NEs, and thus add them to the exclusion lists.

Another procedure for removing inappropriate candidates is by excluding those candidates that contain unlikely N-grams (up to N=5) of characters corresponding to impossible combinations of letters in Bulgarian. We also filter out possible spelling errors by checking for misspelled words (we do up to two letter substitutions and search if the word is in the exclusion lists). For efficiency, we use a list with expected misspellings in Bulgarian (e.g., i-e, o-u, z-s, etc.).

3.5. Grouping of Candidates by means of Stemming

A basic stemmer was implemented in order to reduce the number of candidates by grouping inflected forms. The stemmer matches words longer than 4 letters which share a substring whose length is at least 70\% of each word's length. The forms that meet this requirement are grouped together and the shortest one is assigned the value 'lemma' (simple heuristics for word endings is used to distinguish between forms with equal length). Inevitably, this approach also groups derivationally related words.

3.6. Filtering and Ranking of Candidates

Filtering and ranking of candidates combines metadata information from the corpus samples (namely, date of publication/registration, source, and domain) with statistical analysis.

The candidates are ranked based on frequency and a dispersion measures which account for distribution of the candidate with respect to: (a) documents; (b) sources; (c) domains; and (d) period of time. For each of these features, we divide the documents into text sets (for (a) each text set contains a single document; for (b) we combine all texts from a source; for (c) we combine documents from a domain; for (d)

sets contain documents published in the same period) and calculate the measure DP for any candidate as proposed by Gries (2008). Thus, we have four different measures and rank the candidates based on any of them. In addition, one can apply rules for overall ranking (e.g., by taking average or weighted average). The changes in scores over time can be used to create a timeline of the way a word is established in the language.

3.7. Evaluation

The initial evaluation is based on qualitative rather than quantitative criteria and is performed manually. The results show that the filtering and ranking are of significant importance to render the task of manual validation possible since it considerably reduces the number of candidates (over 20 times).

4. A Web-based System for Observation of Neologisms

The web-based system for observation of neologisms is part of a complex system for collection and analysis of media content in Bulgarian which currently covers texts collection and their linguistic annotation and analysis. The processed text material is then used in various web-based systems: media monitoring (e.g., quotation extraction and attribution – http://dcl.bas.bg/quotations/) and lexicographic work (e.g., neologism detection).

The web-based system for observation of neologism targets the identification of the most probable candidates for neologisms and provides the following functionalities for experienced lexicographers: 1) tracking empirical evidences for potential new words providing data for all registered occurrences over the observed period: the time of appearance and / or the time of registration, if different; the source; the author, if available; and a narrow context; 2) manual evaluation of potential new words which may result in a) removal of non-words, or b) selection of neologisms for inclusion in a dictionary. The evaluation may require observations over a relatively lengthy period of time if there is insufficient evidences for continuous use. The results are available at http://dcl.bas.bg/neologisms/.

The workflow for the collection and retrieval of candidates for new words includes the following components:

(1) Download of texts from several news agencies. Two approaches were implemented: (a) monitoring of RSS feeds and downloading of web pages (limited number of pages, top news only); and (b) focused crawling of selected media sources (involves pre-crawl data mining to optimise the crawling). Extensive metadata are extracted from the original webpage and stored separately from the text. Author's name and date of publication are extracted from the html markup, and the domain is assigned through keyword extraction and analysis. Texts are also applied preprocessing – removal of boilerplate, pictures, etc. This is part of the process of extending and enriching the BulNC with new texts.

(2) Linguistic annotation of newly collected texts as described in Subsection 3.2.

(3) Neologisms retrieval as described in Section 3.

(4) Presentation of results. The results are represented online in a structured manner and with a search functionality.

(5) Update routine. Results are automatically updated on regular intervals throughout the year after newly downloaded data have been processed.

Retrieved candidates for new words are presented online. They are put into a database and allow filtering based on: (a) a character or a sequence of characters (possibly the whole new word); (b) period of time; (c) media; and (d) combinations of (a) – (c).

5. Conclusion and Future Work

In this paper we propose a method for automatic neologism detection integrated into a web-based system for observation of neologisms, which is part of a complex system for collection and analysis of media content.

Our future work on neologism detection will be focused on perfecting the system by introducing more advanced statistical analysis, pattern matching techniques and refined spelling error detection; enhancing the recognition of MWEs; and developing methods for identification of semantic neologisms.

6. References

Falk, I., Bernhard, D. & Gérard, C. (2014). From Non-Word to New Word: Automatically Identifying Neologisms in French Newspapers. In *Proceedings from LREC 2014*, pp. 4337–4344.

Gries, S. (2008). Dispersions and Adjusted Frequencies in Corpora. *International Journal of Computational Linguistics*, 13(4), pp. 403–437.

Kerremans, D., Stegmayr, S. & Schmid H.-J. (2012). The NeoCrawler: Identifying and Retrieving Neologisms from the Internet and Monitoring On-going Change. In K. Allan & J. A. Robinson (eds.) *Current Methods in Historical Semantics*. Berlin: de Gruyter Mouton, pp. 59–96.

Kilgarriff, A., Herman, O., Bušta, J., Rychlý, P. & Jakubíček, M. (2015). DIACRAN: a Framework for Diachronic Analysis. In F. Formato & A. Hardie (eds.) *Corpus Linguistics 2015*. Lancaster: UCREL, pp. 65–70.

Koeva, S. & Genov, A. (2011). Bulgarian Language Processing Chain. In *Proceedings of Integration of Multilingual Resources and Tools in Web Applications. Workshop in conjunction with GSCL 2011, 26 September 2011, University of Hamburg*.

Koeva, S., Stoyanova, I., Leseva, S., Dimitrova, T., Dekova, R. & Tarpomanova E. (2012). The Bulgarian National Corpus: Theory and Practice in Corpus Design.

Journal of Language Modelling, 0(1), pp. 65–110.

O'Donovan, R. & O'Neil, M. (2008). A Systematic Approach to the Selection of Neologisms for Inclusion in a Large Monolingual Dictionary. In *Proceedings of the 13th Euralex International Congress, Barcelona, Spain*, pp. 571–579.

Paryzek, P. (2008). Comparison of Selected Methods for the Retrieval of Neologisms. *Investigationes Linguisticae*, 16, pp. 163–181.

Renouf, A. (1993). Making Sense of Text: Automated Approaches to Meaning Extraction. In *Proceedings of 17th International Online Information Meeting, 7-9 December 1993*, pp. 77–86.

Stenetorp, P. (2010). *Automated Extraction of Swedish Neologisms using a Temporally Annotated Corpus.* Master of Science in Engineering (MSc Eng) Thesis, Royal Institute of Technology (KTH), Stockholm, Sweden. Available at: http://pontus.stenetorp.se/res/pdf/stenetorp2010automated.pdf.