



PD Dr.-Ing. Tino Haderlein
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
Lehrstuhl für Mustererkennung (Pattern Recognition Lab)
Martensstraße 3
91058 Erlangen
Germany

March 25, 2015

Opponent's Review for the Ph.D. Thesis of Ing. Tomáš Brychcín "Distributional Semantics in Language Modeling", West Bohemian University in Pilsen

a) Meaning of the Thesis for the Field

Language models are important for applications using automatic speech processing in order to improve recognition accuracy. Since linguistic approaches require an expert in the respective field, unsupervised methods for the creation of such models are desirable. Especially for highly inflectional languages, such as Czech, the sparse data problem is an essential issue. Mr. Brychcín's thesis contributes successfully to the solution of this problem by introducing a new type of stemmer and, in general, the application of distributional semantics to language modeling. By the combination of three different sources of information, i.e. the stemmer, semantic spaces, and Latent Dirichlet Allocation (LDA), significant improvement of the modeling results has been achieved.

b) Method of Problem Solving, Used Methods and Fulfillment of Targets

The main goal was the introduction of new unsupervised methods for improving the performance of language models with special emphasis on inflectional languages. Three goals were mentioned. The first one was about the analysis of the relationship between morphology and language modeling in inflectional languages. This was fulfilled in the chapter and article about the High Performance Stemmer (HPS). It groups words that are lexically similar and uses these clusters as training data for a maximum entropy classifier. The second goal was the use of semantic information for language modeling with unsupervised training methods. Different types of semantic spaces were tested on inflectional and non-inflectional languages. The created language models showed significantly lower perplexity than the baseline 4-gram model. The third goal was the analysis of out-of-vocabulary (OOV) words. Chapter 5.5 gives a brief theoretical summary on the influence of OOV words on the different approaches presented in the thesis. A systematic analysis is not given in detail, but this would have required intensive work on the available speech corpora, and this would have probably been beyond the scope of the thesis.

c) Results of the Thesis

The results given in all three publications, which are summarized in the thesis, outperform state-of-the-art approaches. Both on inflection removal, and on information retrieval, the High Performance Stemmer is in most cases better than competing approaches. This holds for inflectional and non-inflectional languages. The integration of five semantic spaces into language models proved the suitability of the approach in a real-world machine translation task. The integration of all sources of knowledge, as introduced in the third article (Brychcín and Konopík, Computer Speech and Language, 2015), showed significantly improved BLEU scores in translation tasks.

d) Systematics, Clarity, Formal Elaboration, and Language Level

The thesis is composed of three published articles, an introduction to these articles, and an overview of the topic in general. All parts show a clear structure and writing, the mathematical notation is also clear. The use of English is adequate with very few minor mistakes. The expression 'bag-of-words' is misspelled as 'bag-of-word' in a few cases in Chapter 3. In the Bibliography section, there are a few minor issues, such as missing page numbers.

e) Publications of the Author

The publication list contains four articles published in international journals by Elsevier in 2014 and 2015. Two of them are current publications in the renowned journal *Computer Speech and Language (CSL)*, which is an official publication of the International Speech Communication Association (ISCA). For three articles, Mr. Brychcín is the first author, the only other author is his supervisor, Mr. Konopík. Additionally, the list contains eight conference papers from 2008 to 2014. Mr. Brychcín is the first author of four of them; all papers have one or two co-authors.

f) Recommendation for the Acceptance of the Thesis

The thesis and the publication list show that Mr. Brychcín is able to perform research independently in a structured and very thorough way. For this reason, I recommend the acceptance of his thesis for the granting of the academic title Ph.D.

Questions for the Defense of the Thesis

- How does the High Performance Stemmer (HPS) perform on very short Czech words?
- Does the stemmer know about specific rules, e.g. the ‘volatile e’ (as in *lev* → *lva*)?
- You used a maximum entropy classifier in HPS. Did you consider to use other classifiers? Would a similarity measure based on the Levenshtein distance make sense for the task?
- In the article about the HPS, you write: “It is expected that a word stem is related to the initial part of the word.” Do you regard Czech verbs, such as “telefonovat” and “zatelefonovat”, to have two different meanings because of the different initial parts?
- You further write in that article: “The motivation for this feature is the assumption that the length of the suffixes depends on the length of the stems.” Where does such an assumption come from. Is it reasonable?
- The HPS training corpus for Czech contained texts about different topics. One of them was “international”. Did it contain many out-of-vocabulary words then? Were there maybe more than in the corpora of the other languages?
- In your article “Latent semantics in language models”, you write: “We use 50 buckets in linear interpolation for combining our language models.” Why did you choose that number?
- You further write: “We generate and re-score 5000 hypotheses for each translated sentence.” Are they all different in reality? How fast does their probability drop towards zero?
- In the same article, you present results on the linear interpolation with all sub-models. Why did you not use bucketed linear interpolation as in the other experiments?
- How is the work of Pucher et al. related to your topic?
Michael Pucher, Yan Huang, Özgür Çetin, Combination of latent semantic analysis based language models for meeting recognition. In Proc. 2nd IASTED Int. Conf. on Computational Intelligence (CI 2006), p. 465–469, San Francisco, USA, 2006.
Michael Pucher, Yan Huang, Özgür Çetin, Optimization of latent semantic analysis based language model interpolation for meeting recognition. In Proc. 5th Slovenian and 1st Int. Language Technologies Conf., p. 74–78, Ljubljana, Slovenia, 2006.



Tino Haderlein

Ilya Oparin, Ph.D.

Paris, France

23 Mar 2015

Doctoral Thesis Review

Candidate: Ing. Tomáš Brychcín

Title: Distributional Semantics in Language Modeling

DISCLAIMER: Views and opinions expressed in this review are those of the author and not necessarily of his current employer.

This doctoral thesis contributes to improving language modeling and more generally natural language processing by the use of semantic information in the unsupervised way.

The goals and objectives formulated in the thesis are successfully fulfilled. The main contributions of the thesis (as thesis by publications) are as follows:

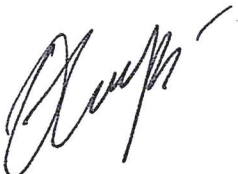
- The author introduces a new unsupervised stemming algorithm - High Precision Stemmer - that is shown to be state-of-the-art as compared to the existing baselines for inflectional languages.
- The author investigates the use of different local-context semantic spaces to improve language modeling baselines with better class-based language models, especially for inflective languages. He introduces the local context semantic space to incorporate the information about word similarity into language models. Not only perplexity results are reported in the experimental section but also machine translation BLEU scores are investigated, with improvements observed for both.
- Finally, the author extends the work on the application of semantic spaces in language modeling by adding long-range semantic dependencies. The combination of three types of approaches: local semantics, global semantics and morphology was studied and positive results were confirmed with perplexity and BLEU score improvements. Such kind of study, that uses all three approaches can be considered novel as it has not been done to this extent before. It was shown that different kinds of information bring additive gains over the baselines.

The scientific approach T.Brychcín pursues to fulfil the goals of the thesis is systematic and correct in the way theoretical foundations are presented, experimental data are chosen, baselines are defined and improvements are reported. The only remark I may have is that if relative improvements are reported, they should always be accompanied by absolute numbers. E.g. relative perplexity improvements cited on page 32 are not fully informative, as one can have the same relative reduction in perplexity with reducing it from 1000 to 900 and from 100 to 90 - though the relative reduction in entropy will be different. If only a relative reduction is reported, it is preferable to use entropy rather than perplexity. Having said that, it should be noted that absolute numbers are sometimes omitted in the introductory or conclusion sections but can always be found in the experimental parts. This makes the experimental results convincing and the above mentioned minor remarks are thus rather concerned with the form of presentation. At the same time, the form of presentation is generally clear, easy to follow and precise.

The criticism I have only relates to minor points. E.g. in several places the author mentions that the distinctive feature of the Kneser-Ney smoothing is about the way unigram probability distribution is calculated. Actually, it can be formulated more generally, as this statement is applied to lower-order probability distributions. Such remarks deal with minor points that do not substantially influence the way the work is presented and they should not be taken into consideration when judging the quality of thesis on the whole. The doctoral thesis of T.Brychcín is a significant work in the domain of natural language processing.

Publication record of T.Brychcín is solid. It includes both papers published in conference proceedings and scientific journals highly rated in the domain of natural language processing.

I recommend the thesis *Distributional Semantics in Language Modeling* for defence and believe T.Brychcín merits receiving the doctoral degree.



Ilya Oparin