Building A Thesaurus Using LDA-Frames

Jiří Materna

Centre for Natural Language Processing
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00, Brno, Czech Republic
xmaterna@fi.muni.cz

Abstract. In this paper we present a new method for measuring semantic relatedness of lexical units, which can be used to generate a thesaurus automatically. The method is based on a comparison of probability distributions of semantic frames generated using the LDA-frames algorithm. The idea is evaluated by measuring the overlap of WordNet synsets and generated semantic clusters. The results show that the method outperforms another automatic approach used in the Sketch Engine project.

Key words: thesaurus, LDA-frames, semantic relatedness, lexical semantics

1 Introduction

Identifying meaning of words is one of the crucial problems in linguistics. While ordinary monolingual dictionaries index words alphabetically and provide a definition for every record, thesauri index words senses and group words with similar meaning. However, there is an important difference between lexicons of synonyms and thesauri. The clustered words in thesauri are not exactly synonyms. Thesauri rather group words with similar patterns of usage or semantically related words across parts of speech. As in other areas of linguistics, there is an important issue of polysemy. Since most of words in natural languages may have different meanings depending on the contexts in which they are used, a word usually belongs to multiple clusters. For instance, the word *bank* has two meanings – a financial institution and a border of a river, thus it should belong into two clusters.

Thesaurus is not only a useful resource helping to find and understand related words or phrases, which is mainly used by writers when hesitating what word they should choose. A word cluster or ranked list of similar words has many applications in natural language processing. One such application is the information retrieval task. In an information retrieval system, the query can be augmented by semantically related terms, which may lead to better retrieval quality.

One of the most popular English thesauri is Roget's thesaurus. It is a widely used English language thesaurus created by Dr. Peter Mark Roget in nineteenth century [6]. Another manually created resource grouping similar

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2012, pp. 97–103, 2012. © Tribun EU 2012

words together is WordNet [3]. WordNet is an electronic lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms called synsets, each expressing a distinct concept. Moreover, synsets are interlinked by means of conceptual-semantic and lexical relations. In comparison to Roget's thesaurus, which is primarily intended to be used by humans, WordNet is more often utilized in natural language processing taks.

Since the manual creation of thesauri, and the dictionaries in general, is a very time-consuming work, there are some attempts to create thesauri automatically by processing corpora. The similarity between words is usually measured by looking at their usage in texts. The same approach is used in a thesaurus generated using the Sketch Engine [7]. The similarity of lexical units in the Sketch Engine is measured by comparing so called word sketches. Word sketches are automatic, corpus-based summaries of a word's grammatical and collocational behaviour, which takes as input a corpus of any language and corresponding grammar patterns. The resulting summaries are produced in the form of a ranked list of common word realizations for a grammatical relation of a given target word.

In this work we proposed a similar method, which, instead of comparing word sketches, compares semantic frames of target words. Because the LDA-frames approach provides a probabilistic distribution over all frames, and is able to distinguish between different word senses, this method acquires better results than the Sketch Engine. It is demonstrated by measuring overlap with WordNet synsets.

2 LDA-frames

LDA-frames [8] is an unsupervised approach to identifying semantic frames from semantically unlabelled text corpora. There are many frame formalisms but most of them suffer from the problem that all frames must be created manually and the set of semantic roles must be predefined. The LDA-Frames approach, based on the Latent Dirichlet Allocation [1], avoids both these problems by employing statistics on a syntactically tagged corpus. The only information that must be given is a number of semantic frames and a number of semantic roles to be identified. This limitation, however, can be avoided by automatic estimation of both these parameters.

In the LDA-frames, a frame is represented as a tuple of semantic roles, each of them connected with a grammatical relation i.e. subject, object, modifier, etc. These frames are related to a predicate via a probability distribution. Every semantic role is represented as a probability distribution over its realizations.

The method of automatic identification of semantic frames is based on the probabilistic generative process. We treat each grammatical relation realization as generated from a given semantic frame according to the word distribution of the corresponding semantic role. Supposing the number of frames is given by the parameter F and the number of semantic roles by R. The realizations are generated by the LDA-Frames algorithm as follows.

For each lexical unit $u \in \{1, 2, \ldots, U\}$:

1. Choose a frame distribution φ_u from $\mathrm{Dir}(\alpha)$.
2. For each lexical unit realization $t \in \{1, 2, \ldots, T\}$ choose a frame f_{ut} from $\mathrm{Mult}(\varphi_u)$, where $f_{ut} \in \{1, 2, \ldots, F\}$:
3. For each slot $s \in \{1, 2, \ldots, S\}$ of the frame f_{ut}
   (a) look up the corresponding semantic role r_{uts} from $\rho_{f_{ut}s}$, where $r_{uts} \in \{1, 2, \ldots, R\}$.
   (b) generate a grammatical realization w_{uts} from $\mathrm{Multinomial}(\theta_{r_{uts}})$

The graphical model for LDA-Frames is shown in the figure 1. In this model, $\rho_{f,s}$ is a projection $(f, s) \mapsto r$, which assigns a semantic role for each slot s of a frame f. This projection is global for all lexical units. The multinomial distribution of words, symbolized by θ_r, for a semantic role r is generated from $\mathrm{Dirichlet}(\beta)$. The model is parametrized by hyperparameters of prior distributions α and β, usually set by hand to a value between 0.01 – 0.1.

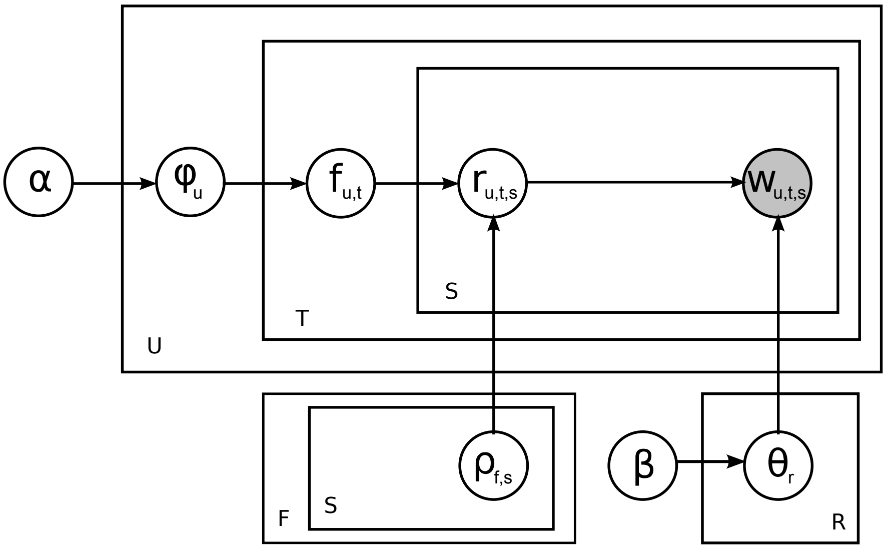

Fig. 1. Graphical model for LDA-Frames.

For the inference we use collapsed Gibbs sampling, where the θ, ρ and φ distributions are marginalized out. After having all topic variables **f** and **r** inferred, we can proceed to computing the lexical unit–frame distribution and the semantic role–word distribution. Let C^{φ}_{uf} be the count of cases where frame f is assigned to lexical unit u, C^{θ}_{rw} is the count of cases where word w is assigned to semantic role r and V is the size of vocabulary. The φ and θ distributions are computed using the following formulas:

$$\varphi_u = \frac{C^{\varphi}_{uf} + \alpha}{\sum_f C^{\varphi}_{uf} + F\alpha} \tag{1}$$

$$\theta_r = \frac{C^{\theta}_{rw} + \beta}{\sum_w C^{\theta}_{rw} + V\beta} \tag{2}$$

3 Measuring Semantic Relatedness

The semantic frames generated by the LDA-Frames algorithm are an interesting source of information about selectional preferences, but they can even be used for grouping semantically related lexical units. Separated semantic frames can hardly capture the whole semantic information about a lexical unit. Nevertheless, the LDA-Frames provide an information about the relatedness to every semantic frame we have inferred. After the inference process, each lexical unit u is connected with a probability distribution over semantic frames φ_u. Thus, we can group lexical units with similar probability distributions together to make a semantic cluster. In this work we will call these clusters *similarity sets*.

There are several methods how to compare probability distributions. We use the Hellinger Distance, which measures the divergence of two probability distributions, and is a symmetric modification of the Kullback-Leibner divergence [5]. For two probability distributions φ_a, φ_b, where $P(f|x)$ is the probability of frame f being generated by lexical unit x, the Hellinger Distance is defined as follows:

$$H(a,b) = \sqrt{\frac{1}{2}\sum_{f=1}^{F}\left(\sqrt{P(f|a)} - \sqrt{P(f|b)}\right)^2} \tag{3}$$

By using the Hellinger distance, we can generate a ranked list of semantically similar words for every lexical unit u the semantic frames have been computed for. Then the similarity set is chosen by selecting n best candidates or by selecting all candidates c, where $H(u,c) < \tau$ for some threshold τ.

4 The Experiment

The experiments have been performed for all transitive verbs having their lemma frequency in British National Corpus grater than 100. The transitiveness has been determined by selecting those verbs that have both subject valency and direct object valency presented in the Corpus Pattern Analysis lexicon [4]. Such constraints have been fulfilled by 4053 English verbs.

In order to generate LDA-frames for those English verbs, we have syntactically annotated British National Corpus using the Stanford Parser [2]. It has provided a set of approximately 1.4 millions of (verb, subject, object) tuples that have been used as the training data for the LDA-frames algorithm. Based on previous experiments [8], we set the number of frames to 1200, number of roles to 400, and the hyperparameters as follows $\alpha = 0.1$, $\beta = 0.1$. After inferring φ distributions for every lexical unit, we have created a list of similar verbs sorted in ascending ordered based on the distance measure described in the previous section. The verbs with distance 1.0 were omitted. Using those data we have created 20 different thesauri. For $1 \leq n \leq 20$ the thesaurus has been built as the set of at most first n items from the similarity lists for every verb.

We have evaluated the quality of the thesauri built using LDA-frames by comparing them to the thesauri obtained from the Sketch Engine. To be fair, the word sketches have been generated on the British National Corpus just using the `subject_of` and `object_of` grammatical relations. The resulting thesaurus is in the form of sorted list of similar words, so we have been able to use the same method as in the case of the LDA-frames thesaurus and to create 20 thesauri in the same way.

It is obvious that not all verbs have got exactly n similar verbs in their similarity sets, because verbs with distance 1.0 were omitted. Table 1 shows average number of words in the similarity sets for every n we considered.

Table 1. Average number of words in the similarity sets.

n	1	2	3	4	5	6	7	8	9	10
LDA-frames	1.0	1.99	2.99	3.99	4.98	5.98	6.97	7.96	8.95	9.93
Sketch Engine	1.0	1.98	2.97	3.94	4.90	5.86	6.80	7.74	8.67	9.59
n	11	12	13	14	15	16	17	18	19	20
LDA-frames	10.91	11.89	12.87	13.85	14.82	15.79	16.76	17.73	18.69	19.66
Sketch Engine	10.50	11.40	12.30	13.18	14.05	14.92	15.77	16.61	17.45	18.27

The results from the table can be interpreted in the way that the Sketch Engine thesauri are stricter than LDA-frames ones and produce smaller similarity sets.

In order to measure the quality of the generated thesauri we have compared the similarity sets with synsets from English WordNet 3.0. First, we have selected verbs contained in both WordNet and our set of verbs. There were 2880 verbs in the intersection. The quality has been measured as the average number of verbs from a similarity set contained in the corresponding WordNet synset, normalized by the size of the similarity set. Formally, let V be the number of investigated verbs v_i, $T(v)$ similarity set for verb v in thesaurus T and $W(v)$ synset for verb v in WordNet:

$$Score(T) = \frac{1}{V} \sum_{v=1}^{V} \frac{|T(v) \cap W(v)|}{|T(v)|} \tag{4}$$

Resulting scores of both the Sketch Engine thesaurus and the LDA-frames thesaurus for similarity set sizes $1 \leq n \leq 20$ is shown in figure 2. One can see that the LDA-frames thesaurus outperforms the Sketch Engine for any choice of the size of similarity sets. The difference is most noticeable when $n = 1$. This special case measure whether the most similar verb is presented in the corresponding WordNet synset. This condition is satisfied in approximately 9.5 % verbs for LDA-frames and 6.5 % for Sketch Engine. The scores may seem to be quite small but it is important that only subject and object grammatical relations have been taken into consideration when computing the similarity. This means, for instance, that English verbs *eat* and *cook* have very high similarity scores, because they both are used in the same contexts and have completely identical semantic frames. It is straightforward that the algorithm could achieve much better results if there were used more than two grammatical relations. Specifically, verbs *eat* and *cook* could be differentiated, for example, by adding a grammatical relation corresponding with what instrument is used for eating or cooking.

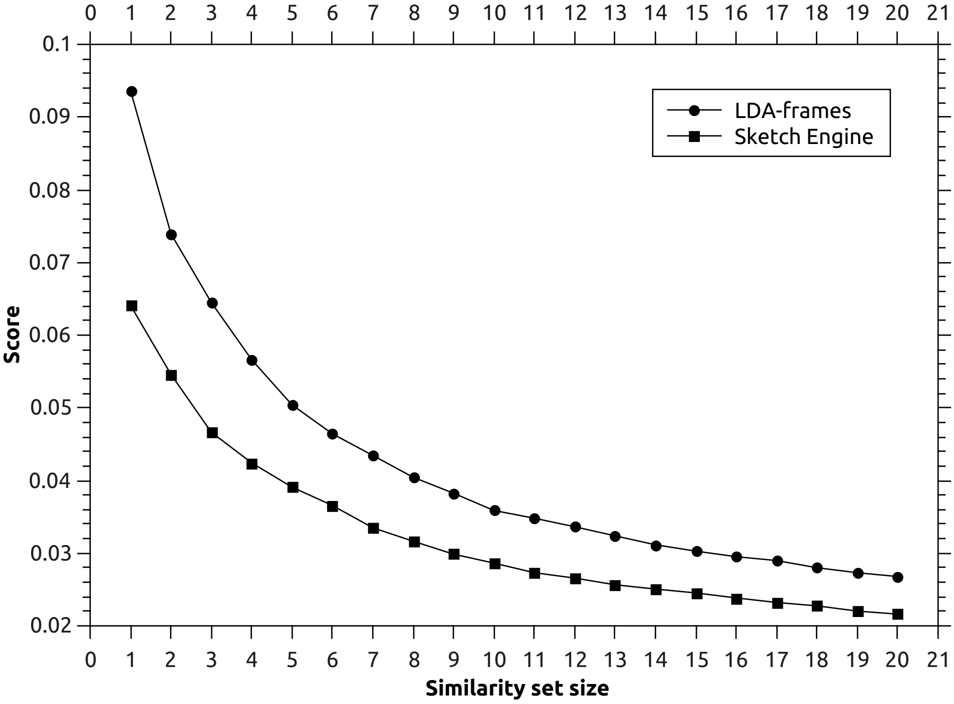

Fig. 2. Comparison of Sketch Engine thesaurus and LDA-frames thesaurus.

5 Conclusion

In this work we presented a new method for automatic building thesaurus from text corpora. The algorithm was applied to texts from the British National Corpus, and the quality was judged by measuring overlap with manually created synsets from WordNet 3.0. The results show that our algorithm outperforms similar approach from the Sketch Engine on the same training data. Only subject and object grammatical relation have been taken into consideration. In the future, we would like to enhance training data by other grammatical relation, which should lead to significantly better results.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013 and by EC FP7 project ICT-248307.

References

1. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res*, 3:993 – 1022.
2. Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating Typed Dependency Parses from Phrase Structure Parses. In *The International Conference on Language Resources and Evaluation (LREC) 2006*, 2006.
3. Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.
4. Patrick Hanks and James Pustejovsky. A Pattern Dictionary for Natural Language Processing. In *Revue Francaise de Langue Appliquée*. Brandeis University, 2005.
5. Michiel Hazewinkel. *Encyclopedia of Mathematics*. Springer, 2001.
6. Werner Hüllen. *A History of Roget's Thesaurus: Origins, Development, and Design*. OUP Oxford, 2003.
7. Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell. The Sketch Engine. In *Proceedings of the Eleventh EURALEX International Congress*, pages 205–116. Lorient, France, 2004.
8. Jiří Materna. LDA-Frames: An Unsupervised Approach to Generating Semantic Frames. In Alexander Gelbukh, editor, *Proceedings of the 13th International Conference CICLing 2012, Part I*, pages 376–387, New Delhi, India, 2012. Springer Berlin / Heidelberg.