Towards clustering-based word sense discrimination

Darja Fišer*, Špela Vintar*, Ljupčo Todorovski†

* Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, SI – 1000 Ljubljana
{spela.vintar, darja.fiser1}@guest.arnes.si
† Fakulteta za upravo
Gosarjeva 5, SI – 1000 Ljubljana
ljupco.todorovski@fu.uni-lj.si

Abstract

This paper describes a series of experiments conducted to group similar words using context features derived from a corpus. The goal is to find an approach that would be suitable for cleaning the fuzzy WordNet synsets obtained by automatic translation of Serbian synsets into Slovene. Similar techniques have been used successfully by a number of researches already and they are attractive particularly because they are knowledge-lean and based on evidence found in simple raw text. A selection of features and settings are tested on sample test sets with an unsupervised machine learning method called hierarchical clustering. In the final part of the paper, the obtained results are analyzed and the optimal set of features is selected, followed by a discussion of the results and some further research plans.

Poskus uporabe hierarhičnega razvrščanja v skupine za določanje pomena besed

Prispevek opisuje niz eksperimentov, s katerimi smo na podlagi okolice besed, ki smo jo izluščili iz korpusa, skušali besede združiti v skupine glede na njihov pomen. Cilj naloge je bil najti pristop, ki bi bil primeren za čiščenje avtomatsko prevedenih sinsetov v slovenskem semantičnem leksikonu. Uporabljene tehnike so pred nami uspešno uporabili že številni avtorji in so priljubljene predvsem zato, ker zanje razen besedilnih zbirk posebni jezikovni viri niso potrebni. V eksperimentih smo na vzorčnih primerih sinsetov preverili različne nize atributov z metodo nenadzorovanega strojnega učenja, imenovanega hierarhično razvrščanje v skupine. Prispevek analizira optimalen niz atributov, predstavlja in vrednoti rezultate razvrščanja in podaja načrte za prihodnost.

1. Introduction

Words in natural language often have multiple distinct meanings which can only be determined by considering the context in which they occur. Given a target word used in a number of different contexts, its senses can be grouped together by determining which contexts are the most similar to each other.

The approach, commonly referred to as word sense discrimination (e.g. Agirre and Edmonds 2006), does not categorize words on a pre-existing sense inventory but clusters words based on their contexts observed from corpora. It is attractive primarily because it is knowledge-lean and thus does not rely on sense-tagged corpora or other manually crafted knowledge resources that are difficult and expensive to obtain. Furthermore, because it is data-driven, it does not fall victim to an absolute view of word meanings encoded in sense inventories (see Kilggarrif 1997, Hanks 2000) and is adaptable and portable across languages.

Word sense discrimination can be carried out either in a mono- or multilingual setting. The distributional approaches make distinctions between word meanings based on the assumption that words which appear in similar contexts have similar meanings (Harris 1968, Miller & Charles 1991). They do not assign but discriminate word meanings based on their distributional similarity found in monolingual corpora. On the other hand, approaches taking advantage of translational equivalence found in word-aligned parallel corpora use the sense-dependent translations of a word as a kind of sense inventory for that word in the source language (Brown et al. 1991, Gale et al. 1992, Ide et al. 2002).

A further distinction between the approaches in the word-sense discrimination domain is whether we are interested in identifying sets of related words by measuring similarity between word co-occurrence vectors (type-based), such as Latent Semantic Analysis (Deerwester et al. 1991), Hyperspace Analogue to Language (Burgess and Lund 2000) and Clustering by Committee (Lin and Pantel 2002). If, however, we aim to distinguish among the senses of a word in multiple contexts by clustering all the contexts of a word, we need to look into token-based approaches (Schütze 1998). In our case, the envisaged application was the construction of the Slovene Wordnet, where one of the tasks includes the validation of automatically translated synsets. It was hoped that the text mining algorithms based on first-order features, word co-occurences and POS would cluster similar words together and help us find the odd ones out. This is why we adopted an approach similar to McQuitty's Similarity Analysis (Pedersen & Bruce 1998).

The paper is organized as follows. In section 2 we briefly describe the process of building the Slovenian Wordnet and give an example of the 'fuzzy' synsets we attempt to clean. Section 3 describes the corpus and the methods used to construct the datasets. Section 4 presents the text mining methods and the distance measures selected for our experiments. Finally, sections 5 and 6 present the results obtained with different settings and discusses them.

2. Building Slovene WordNet

WordNet (Fellbaum 1998) is an extensive lexical database in which words are divided by part of speech and organized into a hierarchy of nodes, where each node represents a concept. Words denoting the same concept

are grouped into a synset, together with links to other relevant synsets (e.g. antonyms).

In recent years, WordNet has become one of the most valuable resources for a wide range of NLP research and applications which initiated the development of WordNets for many other languages (e.g. EuroWordNet[1], BalkaNet[2]). One of such enterprises is the building of Slovene WordNet (see Erjavec & Fišer 2006).

Being limited in the resources and manpower at our disposal, the expand model (Vossen 1998) seemed like the most suitable approach. Synsets were taken from the existing WordNet and were translated into Slovene. We used the Serbian WordNet (SWN) as the closest relative of Slovene in the WordNet family because we believe that concepts and relations among them overlap across languages better if the languages are closely related.

The Jurančič Slovene / Serbo-Croatian bilingual dictionary was inverted to give pairs of Serbo-Croatian / Slovene lemmas. This lexicon was then used to automatically translate Serbian synset literals from Base Concept Sets 1 and 2.

A typical error occurred in translations of polysemous literals where they were translated with equivalents that would be acceptable for some senses but not for this particular one (see Figure 1). Attempts are being made to detect and correct such errors with a clustering technique presented in this paper.

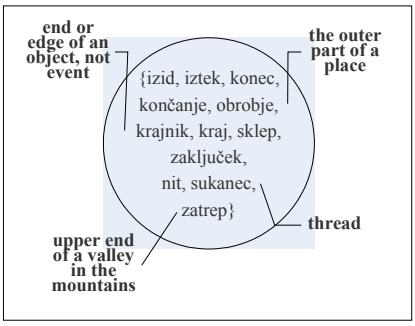

Figure 1. Example of a fuzzy synset: Eng. ending, conclusion, finish (event whose occurrence ends something)

3. Feature and settings selection

3.1. Test sets

As the main aim of our experiments was to test whether our text mining algorithms would prove reliable in grouping together similar words represented by their contexts, we first constructed two controlled Test Sets where each consisted of two distinct groups of synonyms.

- **Test Set 1**: *profesorica*[+], *učiteljica*[+], *tovariš*[+], *tovarišica*[+], *mentor*[+], *učitelj*[+], *profesor*[+]; *veselje*[*], *radost*[*], *sreča*[*], *zadovoljstvo*[*] (Eng. [+]: teacher, [*]: happiness)
- **Test Set 2**: *mož*[+], *fant*[+], *moški*[+], *možak*[+], *deček*[+]; *gora*[*], *hrib*[*], *vzpetina*[*], *grič*[*] (Eng. [+]: man, [*]: mountain)

These two Test Sets were considered "easy" because each consisted of only 2 target clusters, with clearly distinguishable meanings. They were used primarily to define the optimal context features and tune the clustering algorithm to the task at hand.

For a more realistic set of experiments, we adapted some unedited synsets from the Slovenian Wordnet in which the typical polysemy error explained above was observed in order to see whether the clusters proposed by the algorithm would detect the different word senses and whether it could also be used to validate the automatically translated synsets:

- **Test Set 3**: *panoga*[+], *stroka*[+], *disciplina*[+], *veja*[*], *odrastek*[*] (Eng. [+]:branch, division, [*]: tree branch)
- **Test Set 4**: *konec*[+], *kraj*[+], *krajnik*[*], *obrobje*[*], *nit*[*], *sukanec*[*], *zaključek*[+], *sklep*[+], *zatrep*[*] (Eng. [+]:end, conclusion, [*]: other)

3.2. Context features

Each word in a Test Set is described by a number of parameters, where a parameter is defined as a word appearing within the same sentence as the test word, i.e. its context. The parameters and their values were collected from the FidaPlus[3] corpus, a 100-million reference corpus of Slovene (Gorjanc 1999).

A subcorpus was extracted for each dataset in order to speed up the clustering but also because some sort of normalization of the corpus was required. We observed that without a normalization the occurrence of frequent words is disproportional compared to that of infrequent words to the extent that it completely overrides both the selection and distribution of parameters. This is why the same number of sentences for all instances from the datasets were included. If an instance was more frequent than the instance with minimum occurrence, only the number of sentences corresponding to the instance with minimum occurrence were randomly selected and included in the subcorpus.

We were also interested in finding the optimal number of parameters used for clustering. This is why we ran the tests in two different settings. In one we included all the parameters found in the corpus, and in the other we sorted the parameters in the descending order and only included 500 most frequent ones.

Our assumption was that some context features have greater importance for sense discrimination than others. We therefore tested and evaluated several variants of context selection:

- all (lemmatized) tokens within the same sentence (ALL)
- verbs and nouns (VN)
- only verbs (V)
- only nouns (N)
- only adjectives directly preceding the noun (A1)
- only verbs following the noun in question (V1)

The computed parameters were represented either by:

- the Binary Frequency, returning the values 0 or 1 according to the non-occurrence or occurrence of the context word in the corpus (BIN)
- the TFIDF measure, returning a value between 0 and 1 that is computed by multiplying Term Frequency by Inverse Document Frequency

The Inverse Document Frequency (IDF) weighing method is one of the most popular ones in text retrieval

[1] http://www.illc.uva.nl/EuroWordNet/
[2] http://www.ceid.upatras.gr/Balkanet/
[3] http://www.fidaplus.net

methods and language processing techniques (Robertson 2004). It was first proposed by Karen Spärk Jones (1972) and is based on counting the number of documents in the collection which contain the term in question. A query term which occurs in many documents is not a good discriminator and should therefore be given less weight than the one which occurs in few documents. Term Frequency (TF) which is the frequency of a given term in the document itself. In this case, the higher the frequency, the higher the importance of the term in this document.

$$\mathrm{tf} = \frac{n_i}{\sum_k n_k}$$

n_i: no. of occurrences of the considered term
$\sum_k n_k$: no. of occurrences of all terms

$$\mathrm{idf} = \log \frac{|D|}{|(d_i \supset t_i)|}$$

|D|: total no. of documents in the corpus
$|(d_i \supset t_i)|$: no. of documents where the term t_i appears (that is $n_i \neq 0$)

$$\mathrm{tfidf} = \mathrm{tf} \cdot \mathrm{idf}$$

Figure 2. The TFIDF function (adapted from Salton & Buckley 1988)

4. Text mining methods

Word sense discrimination has been a popular topic of interest in the past decade (see Schütze 1998; Pedersen & Bruce 1998). Its core problem is finding classes of similar contexts such that each class represents a single word sense. Contexts that are grouped in the same class represent a particular word sense.

As opposed to related methods, the McQuitty's Similarity Analysis produces a relatively small feature vector of a target word's morphological features, POS of the surrounding words and co-occurence features. A first-order vector is created for each context. These are then compared according to how many features they have in common.

4.1. Hierarchical clustering

Clustering is an unsupervised learning method. Given data about a set of instances, a clustering algorithm creates groups of objects following two criteria. Firstly, instances are close (or similar) to the other instances from the same group (internal cohesion) and secondly, they are distant (or dissimilar) from instances in the other groups (external isolation) (Vintar et al. 2003).

A particular class of clustering methods studied and widely used in statistical data analysis are hierarchical clustering methods. Their main advantage is that the number of clusters does not need to be specified in advance. The agglomerative hierarchical clustering is a bottom-up algorithm that merges clusters into larger and larger units. It starts with assigning each instance to its own cluster, and iteratively joins together the two closest (most similar) clusters. The distances between instances are provided as input to the clustering algorithm. The iteration continues until all instances are clustered into a single cluster (Manning et al. 2006).

The output of the hierarchical clustering algorithm is a hierarchical tree of clusters or dendrogram (see Figure 4) that illustrates the order in which instances are joined together in clusters. Initial clusters, consisting of a single element, form the leaves of the tree and each internal node represents a cluster that is formed by joining its children nodes. The height of the node is proportional to the distance between the joined clusters.

In the final step of the hierarchical clustering algorithm, the dendrogram is cut into sub-trees, producing separate clusters from elements in each sub-tree. Cutting the same dendrogram at different heights produces different number of clusters. The optimal "cut point" that produces clusters with maximal internal cohesiveness and minimal external isolation from a given dendrogram is where the difference between heights of two successive nodes in the tree is maximal (Todorovski et al. 2002).

4.2. Distance measures

4.2.1. Distance measures between data points

For any clustering the choice of measuring the distance between objects and clusters of objects is very important. The most commonly used distance measures are those which define distance between two n-dimensional vectors of real numbers. In the presented experiment, the Manhattan distance measure that captures the difference in the scale and baseline between objects (the sum of distances) was used (see Table 5).

4.2.2. Distance measures between clusters

Variants of the HAC algorithm differ in how similarity is defined. The most widely known distance measures used with hierarchical agglomerative algorithms are single (minimum), average and complete (maximum) links, also known as UPGMA (Purandare & Pedersen 2004). In single-link clustering, distance between two clusters is the distance between the nearest neighbors in those clusters. This single-link merge criterion is local. Attention is paid solely to the area where the two clusters come closest to each other. Other, more distant parts of the cluster and the clusters' overall structure are not taken into account. In complete-link clustering, the distance between two clusters is the distance between the furthest points in those clusters. This complete-link merge criterion is non-local; the entire structure of the clusters is taken into account. In average-link clustering the distance between two clusters is the average of the distances between all the points in those clusters. It is a compromise between the sensitivity of complete-link clustering to outliers and the tendency of single-link clustering to form long chains (Manning et al. 2006).

Shütze (1998) points out that single link clustering tends to place all instances into a single elongated cluster, whereas Purandare (2003) shows that average links produce satisfactory results. In order to determine which function would produce the best result in our setting, we decided to test all three.

5. Results and discussion

This section presents the results for a number of feature configurations, context representations and clustering algorithms. All the Test Settings that were run on each of the four Test Sets from Section 3.1 are listed in Table 1.

Test Settings

normaliz. of subcorpus	yes	no				
no. of parameters	500	unlimited				

context selection	ALL	NV	N	V	V1	A1
weighing method	BIN	TFIDF				
UPGMA link	single	average	complete			

Table 1. Test settings used for clustering

At an early stage of experiments it turned out that words in the Test Sets had highly uneven frequency distributions in the FidaPlus corpus. For example, the lemma *učitelj* occurred 13.241 times and the lemma *tovarišica* 253 times. This meant that overall, the parameters of the frequent lemma had much higher scores than the parameters of the rare one, and even though they shared many of the parameters, they were not treated as similar by the clustering algorithm.

This is why we normalized the subcorpus by including the same number of sentences for each word in a given Test Set, regardless of their number of occurrence in the corpus. First, the word with the least occurrences was identified and all the sentences in which it occurred were included in the subcorpus. As far as the rest of the words from the Test Set are concerned, only the same number (i.e. the number of the most infrequent word in the test set) of randomly selected sentences were included in the subcorpus. This approach yielded much better results, which is why the rest of the Test Settings were compared only on the normalized subcorpus.

The second feature we were interested in was to determine the extent of the context that is the most useful for successful clustering. The first motivation was a practical one; to speed up processing time. The second reason was of a more serious nature; because data sparseness is a well-known problem in clustering, the number of parameters should not be too low. On the other hand, too much noise can have a negative effect on the results as well. Bearing this in mind, we tested the feature with two different settings; once the parameters were sorted in the descending order according to the frequency of their co-occurrence with the words from the Test Sets and only used the 500 most frequent ones, and then all the parameters were used to compute distance measures.

Limiting the number of frequencies turned out to be useless when context selection was very restrictive (i.e. only verbs that appear directly after a given noun from the Test Set) because in most cases, there were less than 500 parameters in the first place. We also found that after having normalized the subcorpus, the datasets became much smaller so that processing time played no role. For example, the total number of parameters in Test Set 2 when all words from all parts of speech were used was 18.309. We believe that any noise brought into the data in this way could be reduced with the appropriate weighing method. We therefore decided not to limit the number parameters in further tests.

Next, we examined what context selection performs best. Our intuition was that different kinds of words co-occurring with the words from the test sets or patterns have varying impact on displaying their semantic (dis)similarity. Since our test sets contain nouns only, we tried to find the best carriers of semantic distance between them.

We repeated tests a number of times, allowing different context features each time: all the words that appear in the same sentence as a given word from the data set (ALL), only nouns and verbs from the same sentence (NV), nouns only (N), verbs only (V), adjectives that directly precede a test word (A1) only and verbs that directly follow a test word only (V1).

Before the corpus was normalized, none of the results were satisfactory but the A1 setting was by far the best. In the normalized corpus the ALL NV and N settings performed much better. The results in the A1 setting did not change while the V and V1 settings turned out to be the worst of all. When examining the distribution of parameters across different POS it becomes clear that good performance of the three settings can be explained by a rich representation of nouns in ALL and NV. In both Test Sets, nouns represent more than half of all the parameters in ALL and almost 80 per cent of all the parameters in NV (see Table 2). This means that the parameters in these two settings do not differ much from the N setting, explaining the similarity of the results obtained by clustering.

Test Set 1 (ALL)

N	8213	53,81%
A	3200	20,96%
V	2153	14,11%
other	1698	11,12%
total	15264	100,00%

Test Set 2 (ALL)

N	9293	50,76%
A	4098	22,38%
V	2540	13,87%
other	2378	12,99%
total	18309	100,00%

Test Set 2 (NV)

N	8240	79,28%
V	2153	20,72%
total	10393	100,00%

Test Set 2 (NV)

N	9327	78,60%
V	2540	21,40%
total	11867	100,00%

Table 2. Parameters across POS for Test Sets 1 and 2

Due to disproportionate frequencies of the words in our Test Sets a weighed representation of parameters found in the corpus was necessary. Two popular weighing methods were used; BIN and TFIDF. The latter performed significantly better in all the tests we ran. It is very interesting that both BIN FREQ. and TFIDF measure resulted in consistent clusters of male/female pairs for *teacher: učitelj/učiteljica, profesor/profesorica, tovariš/tovarišica*. BIN separates the frequent expressions (e.g. *učitelj/učiteljica, profesor/profesorica*) from their less frequent synonyms (e.g. *tovariš/tovarišica*). TFIDF does not repeat the same mistake and treats all the three pairs equally.

A comparison of graphs created based on shared features and Test Settings, with the three UPGMA options being the only distinction, reveals that single link produces the least satisfactory results. As already reported by Schütze (1998), all the single link graphs have a distinct cascade-like structure and are as such useless for our word task. Much better results were obtained from the average and complete links, with the complete performing slightly better in all the cases.

Table 3 shows clustering reports for the two Test Sets, the parameters for which were obtained from the normalized subcorpus and represented with the TFIDF measure. The HAC algorithm used the Manhattan distance measure and complete link. In both cases, 8 clusters were obtained. But this should not be considered as a bad result since the default cut point can be raised to the level which would leave us with two clusters only.

Finally, because we were interested in the capacity to deal with naturally fuzzy synsets, we selected the ones which contain mistakes resulting from polysemous literals in the source language. The two synsets were pre-edited in order to omit literals which do not occur in the FidaPlus corpus or which are very rare (less than 10 occurrences).

Test Set 1 (ALL, NV)
Manhattan DM (complete) - 8 cl.

cl. 1 (2 el.):	učitelj, učiteljica
cl. 2 (2 el.):	professor, profesorica
cl. 3 (2 el.):	tovariš, tovaršica
cl. 4 (1 el.):	mentor
cl. 5 (1 el.):	sreča
cl. 6 (1 el.):	veselje
cl. 7 (1 el.):	radost
cl. 8 (1 el.):	zadovoljstvo

Test Set 2 (ALL, NV)
Manhattan DM (complete) - 8 cl.

cl. 1 (2 el.):	mož, možak
cl. 2 (1 el.):	fant
cl. 3 (1 el.):	moški
cl. 4 (1 el.):	deček
cl. 5 (1 el.):	hrib
cl. 6 (1 el.):	grič
cl. 7 (1 el.):	vzpetina
cl. 8 (1 el.):	gora

Table 3. Clustering reports for Test Sets 1 and 2

The best results for Test Sets 3 and 4 were obtained by using the normalized subcorpus and an unlimited number of parameters. Context features were NA or N, weighed with the TFIDF measure. The Manhattan DM was computed and a complete link used in the algorithm. It is interesting that BIN performed well in these cases too. This combined with the fact that the ALL setting was less accurate with Test Sets 3 and 4, we took a closer look at their parameters (see Table 4). It turns out that (due to low occurrence of the lemma *odrastek* and consequently a small normalized corpus) there are very few parameters (1705). The nouns contributed less to the overall POS distribution in the ALL setting which influenced the final result. The output of the clustering algorithm is a graph of three clusters which could be cut at a slightly higher point, thus creating two correct clusters (see Figure 4).

Test Set 3
POS of parameters

N	727	42,64%
A	327	19,18%
V	213	12,49%
other	287	18,47%
total	1705	100,00%

Manhattan DM (complete) - 3 clusters

cl. 1 (2 el.):	odrastek, veja
cl. 2 (2 el.):	disciplina, stroka
cl. 3 (1 el.):	panoga

Table 4. Results for Test Set 3

6. Conclusions and Future Work

This paper presents a series of experiments aimed at grouping similar words using context features derived from the reference corpus of Slovene language. The motivation was to try to get around the knowledge-acquisition bottleneck by finding a suitable knowledge-lean approach that would be suitable for cleaning the fuzzy WordNet synsets obtained by automatic translation of Serbian synsets into Slovene. A number of features and settings were tested, after which the data was fed into the agglomerative hierarchical clustering algorithm. The analysis of the results has shown that the approach is promising and should be carried out on a larger scale.

All the experiments conducted performed better with normalized subcorpora and the TFIDF measure. Nearly all of them give better results with an unlimited number of parameters and with nominally strong contexts (ALL, NV, N). Although both average and complete links gave similar results, the height difference between sub-clusters was greater in complete links, making it easier to determine the appropriate cut-off point for clusters.

It must be noted here that with smaller quantities of data, there is less possibility of finding words which share the same contexts. They are more often conceptually related than lexically the same. This is why the approach is not as successful in such cases.

The presented approach is only a preliminary feasibility study and therefore only took into consideration nominal synsets. This is why the optimal feature and settings selection might not hold for verbs, adjectives and adverbs. We plan to carry out a comprehensive survey of optimal feature and settings selection for other parts of speech in the future.

Last but not least, an evaluation method of clustering results needs to be employed in order to enable a comprehensive comparison and evaluation of the results. This is not a trivial task as it is known to be challenging to evaluate the results without manually inspecting them or comparing them to a gold standard (e.g. Schütze 1998, Pedersen & Bruce 1998).

7. References

Agirre, E.; Edmonds, P. (2006): *Word Sense Disambiguation*. Algorithms and Applications. Dordrecht: Springer.

Brown, P.; Della Pietra, S. A.; Della Pietra, V. J.; Mercer, R. L. (1991): *Word-sense disambiguation using statistical methods*. In: Proceedings of the 29th Meeting of the Association for Computational Linguistics (ACL), Berkeley, U.S.A.

Burgess, C.; Lund, K. (2000): *The dynamics of meaning in memory*. Cognitive Dynamics: Conceptual Representational Change in Humans and Machines, ed. by Dietrich E.; Markman, A. 117-156. Mahmah, U.S.A.: Lawrence Erlbaum Associates.

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; Harshman, R. (1990): *Indexing By Latent Semantic Analysis*. Journal of the American Society For Information Science, 41:391-407.

Erjavec, T.; Fišer, D. (2006): *Building Slovene WordNet*. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC '06, Genoa, Italy.

Fellbaum, C. (ed.) (1998): *WordNet: An Electronic Lexical Database*. MIT Press.

Gale, W.; Church, K. W.; Yarowsky, D. (1992): *Using bilingual materials to develop word sense disambiguation methods*. In: Proceedings of the 41th International Conference on Theoretical and Methodological Issues on Machine Translation, Montreal, Canada.

Gorjanc, V. (1999): *Korpusi v jezikoslovju in korpus slovenskega jezika FIDA*. In Proceedings of the 35st Seminar of The Slovene Language and Literature. Ljubljana, Slovenia. 47-59.

Gorjanc, V., Logar, N. (2005): *Od splošnih do specializiranih korpusov - načela gradnje glede na njihov namen*. In the Proceedings Razvoj slovenskega strokovnega jezika, Ljubljana, Slovenia, p. 16.

Hanks, P. (2000): *Do word meanings exist?* In Computers and the Humanities, 34(1-2): 205-215.

Harris, Z. (1968): *Mathematical Structures of Language*. New York: Interscience Publishers.

Ide, N.; Erjavec, T.; Tufis, D. (2002): *Sense discrimination with parallel corpora*. Proceedings of the ACL SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, Philadelphia, U.S.A.

Kilgarrif, A. (1997): *I don't believe in word senses*. In Computers and the Humanities, 31(2): 91-113.

Lin, D. (1998). *Automatic retrieval and clustering of similar words*. In the Proceedings of COLING-ACL98, Montreal, Canada.

Lin, D.; Pantel, P.(2002): *Concept discovery from text*. Proceedings of the 19th International Conference on Computational Linguistics (COLING), Taipei, Taiwan 577-583.

Manning, C.; Prabhakar, R.; Schütze, H. (2006): *An introduction to information retrieval*. Draft. Cambridge University Press.

Miller, G. A.; Charles, W. G. (1991): *Contextual correlates of semantic similarity*. Language and Cognitive Processes. 6(1):1-28.

Pedersen T.; Bruce, R. (1998): *Knowledge lean word sense disambiguation*. In Proceedings of the Fifteenth National Conference on Empirical Methods in Natural Language Processing. Providence, RI, pp. 197-207.

Purandare, A (2003): *Discriminating among word senses using McQuitty's similarity analysis*. In Proceedings of the HLT-NAACL 2003 Student Research Workshop, Edmonton, Alberta, Canada. pp. 19-24.

Purandare, A.; Pedersen, T. (2004): *Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces*. In the Proceedings of the Conference on Computational Natural Language Learning (CoNLL), Boston, USA

Robertson, S. (2004): *Understanding Inverse Document Frequency: On theoretical arguments for IDF*. Journal of Documentation, 60:503-520.

Salton, G.; Buckley, C. (1988): *Term-weighting approaches in automatic text retrieval*. Information Processing & Management 24(5):513-523.

Schütze, H. (1998): *Automatic word sense discrimination*. Computational Linguistics. 24(1):97-123.

Spark Jones, K. (1972): *A statistical interpretation of term specificity and its application in retrieval*. Journal of Documentation, 28:11-22.

Todorovski, L.; Cestnik, B.; Kline, M.; Lavrač, N.; Džeroski, S. (2002): *Qualitative Clustering of Short Time-Series: A Case Study of Firms Reputation Data*. In the Proceedings of the ECML/PKDD'02 workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning. Helsinki, Finland. pp 141-149.

Vintar, Š.; Todorovski, L.; Sonntag, D.; Buitelaar, P. (2003) *Evaluating Context Features for Medical Relation Mining*. In Proceedings of the Workshop on Text Mining and Data Mining for Bioinformatics, ECML/PKDD 2003.

Vossen, P. (ed.) (1998): *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Press.

	profesorica	učiteljica	tovariš	tovarišica	mentor	učitelj	profesor	veselje	radost	sreča	zadovoljstvo
profesorica	0	1.1834	1.4745	1.3767	1.0859	1.1252	0.5471	1.4238	1.5401	1.3118	1.3651
učiteljica	1.1834	0	1.4110	1.3103	1.1375	0.4871	1.1099	1.3551	1.4727	1.2742	1.3035
tovariš	1.4745	1.4110	0	0.8018	1.4432	1.3723	1.4585	1.4774	1.6137	1.3838	1.4460
tovarišica	1.3767	1.3103	0.8018	0	1.3625	1.3069	1.4134	1.4149	1.5217	1.3542	1.3832
mentor	1.0858	1.1375	1.4432	1.3625	0	1.0425	1.1046	1.3880	1.5193	1.2347	1.3286
učitelj	1.1252	0.4871	1.3723	1.3069	1.0425	0	0.9754	1.3120	1.4476	1.1995	1.2503
profesor	0.5471	1.1099	1.4585	1.4134	1.1046	0.9754	0	1.4286	1.5554	1.2965	1.3743
veselje	1.4238	1.3551	1.4774	1.4149	1.3880	1.3120	1.429	0	1.2023	1.1484	1.1696
radost	1.5401	1.4727	1.6137	1.5217	1.5193	1.4476	1.5554	1.2023	0	1.2850	1.3671
sreča	1.3118	1.2742	1.3838	1.3542	1.2347	1.1995	1.2965	1.1484	1.2850	0	1.1181
zadovoljstvo	1.3651	1.3035	1.4460	1.3832	1.3286	1.2503	1.3743	1.1696	1.3671	1.1181	0

Table 5. Manhattan distance measures for Test Set 1 (times 10^{-4})

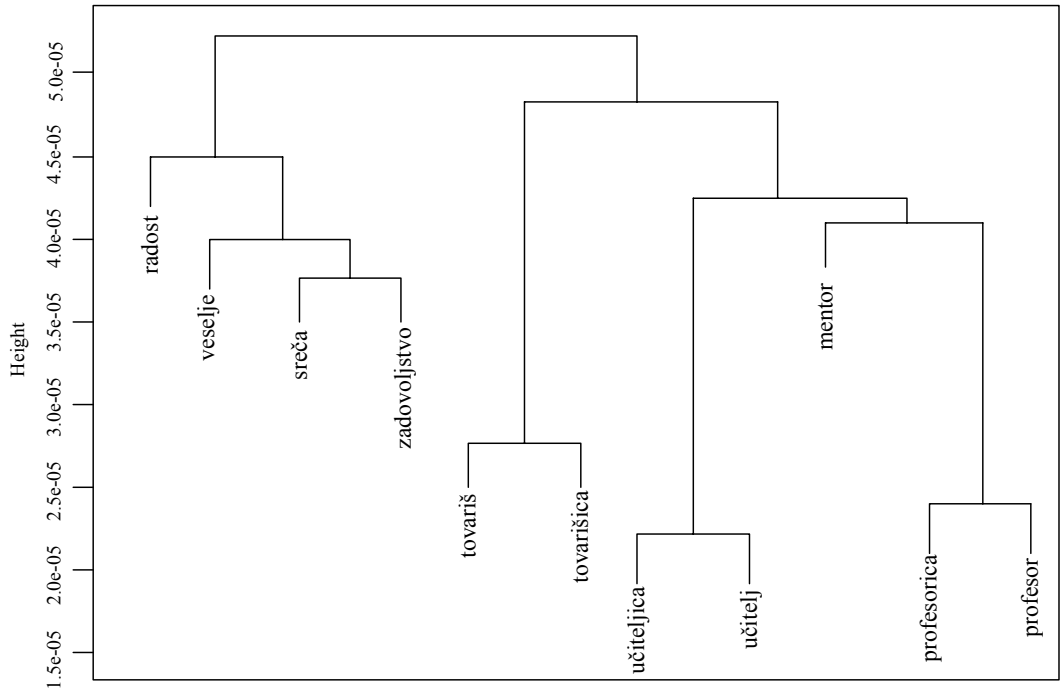

Figure 4. Dendrogram for Test Set 1 (TFIDF, N, complete link)