University of Ljubljana
Faculty of Computer and
Information Science

Cross-lingual sense disambiguation

Matej Miočić, Marko Ivanovski, Matej Kalc

Abstract

Word embeddings are unable to model the property of words that have potentially multiple meanings. Consequently, they cannot capture the particular meaning in which such a word has been used. We will try to tackle this problem for the Slovene language by first constructing our own corpora and then using non-contextual embeddings like Word2Vec and fastText.

Keywords
NLP, sense disambiguation, Slovene homonyms, web-scraping, Word2Vec, fastText, cosine distance

Advisors: Slavko Žitnik

Introduction

Like other languages, the Slovenian language has loads of polysemous words which appear in sentences having different meanings such as ”gol” and ”klop”. Such ambiguity can be a problem for many different NLP tasks resulting in a need for a model capable of understanding in which context the homonym appeared. This project is a simplified version where the main task is to determine whether the word is used in the same sense in both sentences. In this second submission, we will explain how we were able to form our own corpora and then used non-contextual neural embeddings to try and solve this task.

Methods

We divided this task into two parts:

1. Corpus preparation
2. Classification and analyses

Corpus preparation

For the first part of our task, we needed to find as many polysemous words in the Slovenian language as possible and make a list of them. We decided to use a semi-automatic approach that involves web scraping Slovenian’s dictionary web page ”fran.si”. For each Slovenian word, the website contains entries with the meanings of that word. The web scraper was initialized with the 4,768 most frequent Slovenian words. We obtained a list that contains **187** words with multiple meanings. Examples of some words in the list: avgust, bar, draga, faks, gol, klop, metal, rak, servis...

Once the list of words was obtained we proceeded to extract sentences that contain these polysemous words. We decided to use `ccGigafida` [1] which consists of paragraph samples from 31,722 documents. A word in the Slovene language can take multiple forms, such as being in a different case, having a singular, dual, or plural form, and having a different gender. To solve this problem we lemmatized the sentences and obtained the dictionary form of the words by using `classla`. [2]. If a sentence contained a word from our polysemous words list, we extracted the original sentence and its lemmatized form (since some models use non-lemmatized word forms). We extracted a total of **2,511,054** sentences. For classification and analyses, we decided to focus on 5 selected words, which we thought could yield good results. In Figure 1 we show how many sentences we extracted for the selected words.

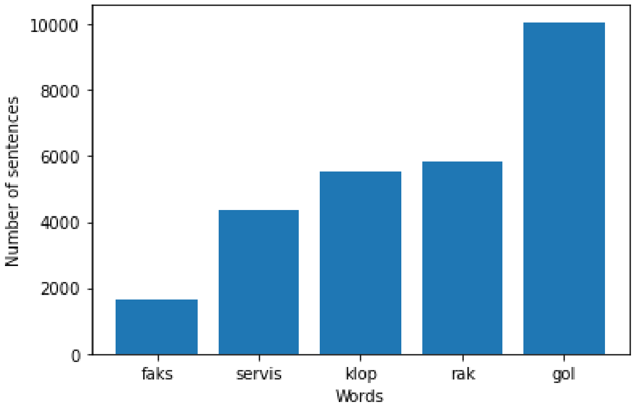


Figure 1. Number of sentences extracted for chosen words.

Classification and analyses

After obtaining the corpus we started off by using neural non-contextual embeddings. First, we removed all duplicate sentences. Since the models we used are non-contextual and static (they don't depend on the context of the sentence) we came up with our own solution on how to represent every sentence into an embedding:

1. Remove the keyword from the sentence.
2. Obtain the vector for the rest of the words in the sentence.
3. Calculate the sentence centroid by averaging the vectors. This is the final embedding for the corresponding sentence.

For obtaining the word vectors we used pre-trained models (in the Slovene language) such as word2vec [3] and fasttext [4]. After obtaining the embedding matrix for all sentences we proceeded by calculating the cosine distance (or similarity) for pairs of these embeddings. These distances were then used in two ways:

- We inspected the pair of sentences which had the best and worst similarity score.
- We also clustered the embeddings for a single word.

During our work we came up with 3 parameters that can be tuned for the need:

- Minimum number of words that a sentence must have to be further analyzed
- Minimum neighbor distance which defines how many words in the neighborhood of the keyword we take into account.
- Minimum similarity score required for a sentence to be used in the clustering.

Results

In the following analysis, we used **8** as the minimum number of words required in a sentence, **6** as the neighborhood distance, and **0.9** as the minimum strong similarity. **Why?** We noticed sentences with very few words do not carry enough information to be compared. Having neighborhood 6 limits only the closest words to influence the embeddings. Also, by applying the minimum similarity requirement we got rid of some isolated sentences which resulted in compromised clustering results.

Best and worst similarity score analysis

The best and worst similarity scores are not bulletproof methods for differentiating sentences. However, we noticed that sentences that have a very high similarity score will definitely have the keyword in the same context. This does not however hold in the opposite case. Some examples for best and worst scores obtained on both models can be seen in Table 1 and 2.

Agglomerative Clustering

Depending on the word we choose the appropriate number of clusters. word2vec and fasttext embeddings were clustered using linkages: single, average, and complete. In the case of ”rak”, word2vec embedding clusters are divided into groups, where each represents a certain meaning, which is the desired result (Table 3). For the same settings, fasttext embedding clusters are not divided into group meanings like word2vec embedding clusters. For the word **gol** word2vec obtained two clusters where one had only football-related sentences are the other one had the naked context but still contained soccer sentences. Similar results were with the words **faks** where one cluster contains the communication context, and the other one is a bit mixed. Klop was the most challenging since one cluster only had sentences about players sitting on a bench. All in all, we concluded that this clustering pretty much depends on how wide or specific the sentences in the context of the words can be. If the sentences in one context are all similar this method returns good results.

Spectral Clustering

When used with word2Vec, spectral clustering was more strict in defining similarities between sentences, hence the small clusters were sometimes smaller when compared with the corresponding agglomerative cluster (Figure 2). However, in some cases, it returned very similar results as the agglomerative clustering with average linkage.
When used with fastText, spectral clustering always returned a cluster containing the majority of sentences and a cluster containing only a few sentences. The output is useless.

DBSCAN Clustering

Instead of defining the number of meanings, we can define the maximum distance between two samples for one to be considered as in the neighborhood of the other using DBSCAN [5]. Unfortunately defining a constant parameter, which clusters all data points into meaning clusters is a real challenge. Tuning the parameter to work well on a single word returns bad results on another one and does not comply with the concept of generalization. The results of DBSCAN are poor, when compared to agglomerative clustering and were omitted from this report.

Conclusion

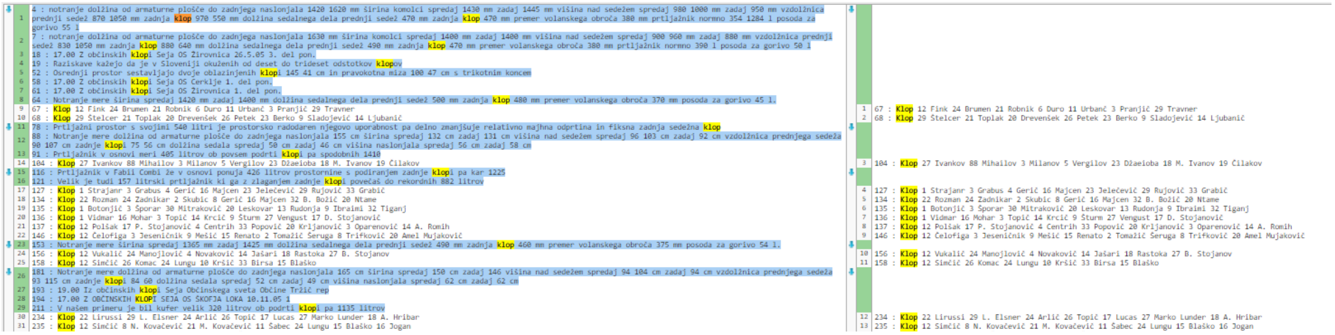
NLP disambiguation tasks aren't trivial, but we have managed to partially solve them. We saw that a very high similarity score results in the same context sentences, but a low score does not necessarily result into the same context clusters. We also saw that clustering results in at least one cluster that contains the word in the same meaning (”rak” and ”golf”). The same is more difficult to achieve with words like ”gol” and ”servis”. One additional difficulty is having a very imbalanced sentence set (with different contexts) for one keyword. When comparing fastText and word2vec, word2vec completely outperformed fastText. In the next part, we will try to use our corpora on contextual embeddings like Elmo and BERT and will compare the results.

Score	Sentence 1	Sentence 2
0.994	**Goli** domači 38 gostje 51 skupaj 89 redni del 73 kazenski streli 16 najboljši strelci ...	**Goli** domači 41 gostje 36 skupaj 77 kartoni rumeni 24 rdeči 4 skupaj 28 najboljši strelci ...
0.988	Odločilni **gol** je v 48. minuti dosegel Oleg Saprikin iz navidez nenevarne akcije	Odločilni **gol** je v 72. minuti dosegel Castillo.
0.984	Edini **gol** je v 66. minuti dosegel Jugoslovan Predrag Mijatović	Edini **gol** je v 59. minuti dosegel Milan Purović rdeče beli so v 45. minuti zastreljali tudi najstrožjo kazen Koroman.
0.220	Takrat se je namreč srečanje končalo brez zadetkov obe moštvi pa sta se osredotočili predvsem na razbijanje nasprotnikovih napadov in branjenje svojega **gola**	Tel. 07 497 5021 Prodam motokultivator frezo **Gol** doni Labin diesel 14 ks s priključki malo rabljen cena po ogledu in do govoru
0.217	V DP je odigral 621 tekem dosegel 332 **golov** in 343 podaj	**Gole** veje brez listja izgledajo kot korenine in Britanci so zanj uporabili izraz upside down od zgoraj navzdol
0.212	Takrat se je namreč srečanje končalo brez zadetkov obe moštvi pa sta se osredotočili predvsem na razbijanje nasprotnikovih napadov in branjenje svojega **gola**	MEDVODE **GOLO** BRDO 170 m2 na parceli 955 m2 l. 1990 mirno naselje možnost treh stanovanj

Table 1. Three most and three least similar sentences using **word2vec** for word **gol**.

Score	Sentence 1	Sentence 2
0.986	Trenutno so najboljši strelci lige Denis Omanovič s 17 Benjamin Arslanovič z 12 Slavko Brečko z 11 in Simon Korun z 10 **goli**	Domači so z izredno obrambo in hitrimi protinapadi navdušili gledalce še posebej pa sta blestela vratar Podpečan z 18 obrambami in Cvijić z 10 **goli**
0.983	Gostujoči vratar Roberto Luongo je zbral kar 48 obramb drugi junak pa je bil finski napadalec Olli Jokinen ki je dosegel dva **gola** natančen pa je bil tudi v dodatnih minutah	Najboljši je bil igralec sredine igrišča Kansas Citya Preki ki je dosegel dva **gola** enkrat pa je bil podajalec
0.978	Zmaji so slavili s 5:0 vse **gole** so dosegli v prvem polčasu	Aris je zmagal s 4:2 gostitelji so vse **gole** dosegli v prvem polčasu
-0.302	Z **goli** se ne obremenjujem če bo priložnost bom seveda poskušal zadeti drugače pa je pri prostih strelih ti so še vedno domeni Zahoviča	Največ doseženih **golov** SCT Olimpija 502 Maribor 359 Hit Gorica 309 največ zmag SCT Olimpija 125 Maribor 111 Mura 96 za tri točke Hit Gorica 36 Maribor 35 Primorje 32 največ neodločenih izidov Hit Gorica 65 ...
-0.317	Največ doseženih **golov** SCT Olimpija 502 Maribor 359 Hit Gorica 309 največ zmag SCT Olimpija 125 Maribor 111 Mura 96 za tri točke Hit Gorica 36 Maribor 35 Primorje 32 največ neodločenih izidov Hit Gorica 65 ...	Trener Kasim Kamenica pa vendarle ni mogel mimo nedeljske tekme z Montpellierjem na kateri je njegova ekipa slavila zmago a ne z razliko v **golih** ki jo je želela
-0.321	**Goli** domači 261 gostje 180 skupaj 441 najboljši strelci Rok Mordej 15 Drobne Goranovič Boštjan Uršič 12 Osredkar Vojsk Vrhovec 11 Drobnič Pertič Stres 10 ...	Z **goli** se ne obremenjujem če bo priložnost bom seveda poskušal zadeti drugače pa je pri prostih strelih ti so še vedno domeni Zahoviča

Table 2. Three most and three least similar sentences using **fastText** for word **gol**.



```
4 : notranje dolžina od armaturne plošče do zadnjega naslonjala 1420 1620 mm širina komolci spredaj 1430 mm zadaj 1445 mm višina nad sedežem spredaj 980 1000 mm zadaj 950 mm vzdolžnica prednji sedež 870 1050 mm zadnja klop 970 550 mm dolžina sedalnega dela prednji sedež 470 mm zadnja klop 470 mm premer volanskega obroča 380 mm prtljažnik normno 354 1284 l posoda za gorivo 55 l
7 : notranje dolžina od armaturne plošče do zadnjega naslonjala 1630 mm širina komolci spredaj 1400 mm zadaj 1400 mm višina nad sedežem spredaj 900 960 mm zadaj 880 mm vzdolžnica prednji sedež 830 1050 mm zadnja klop 880 640 mm dolžina sedalnega dela prednji sedež 490 mm zadnja klop 470 mm premer volanskega obroča 380 mm prtljažnik normno 390 l posoda za gorivo 50 l
18 : 17.00 Z občinskih klopi Seja OS Žirovnica 26.5.05 3. del pon.
19 : Raziskave kažejo da je v Sloveniji okuženih od deset do trideset odstotkov klopov
52 : Osrednji prostor sestavljajo dvoje oblazinjenih klopi 145 41 cm in pravokotna miza 100 47 cm s trikotnim koncem
58 : 17.00 Z občinskih klopi Seja OS Cerklje 1. del pon.
61 : 17.00 Z občinskih klopi Seja OS Žirovnica 1. del pon.
64 : Notranje mere širina spredaj 1420 mm zadaj 1400 mm dolžina sedalnega dela prednji sedež 500 mm zadnja klop 480 mm premer volanskega obroča 370 mm posoda za gorivo 45 l.
67 : Klop 12 Fink 24 Brumen 21 Robnik 6 Duro 11 Urbanč 3 Pranjić 29 Travner
68 : Klop 29 Štelcer 21 Toplak 20 Drevenšek 26 Petek 23 Berko 9 Sladojević 14 Ljubanić
78 : Prtljažni prostor s svojimi 540 litri je prostorsko radodaren njegovo uporabnost pa delno zmanjšuje relativno majhna odprtina in fiksna zadnja sedežna klop
88 : Notranje mere dolžina od armaturne plošče do zadnjega naslonjala 155 cm širina spredaj 132 cm zadaj 131 cm višina nad sedežem spredaj 96 103 cm zadaj 92 cm vzdolžnica prednjega sedeža 90 107 cm zadnje klopi 75 56 cm dolžina sedala spredaj 50 cm zadaj 46 cm višina naslonjala spredaj 56 cm zadaj 58 cm
91 : Prtljažnik v osnovi meri 405 litrov ob povsem podrti klopi pa spodobnih 1410
104 : Klop 27 Ivankov 88 Mihailov 3 Milanov 5 Vergilov 23 Džaeioba 18 M. Ivanov 19 Čilakov
116 : Prtljažnik v Fabii Combi že v osnovi ponuja 426 litrov prostornine s podiranjem zadnje klopi pa kar 1225
121 : Velik je tudi 157 litrski prtljažnik ki ga z zlaganjem zadnje klopi povečaš do rekordnih 882 litrov
127 : Klop 1 Strajanr 3 Grabus 4 Gerić 16 Majcen 23 Jelečević 29 Rujović 33 Grabić
134 : Klop 22 Rozman 24 Zadnikar 2 Skubic 8 Gerić 16 Majcen 32 B. Božić 20 Ntame
135 : Klop 1 Botonjić 3 Šporar 30 Mitraković 20 Leskovar 13 Rudonja 9 Ibraimi 32 Tiganj
136 : Klop 1 Vidmar 16 Mohar 3 Topić 14 Krcić 9 Šturm 27 Vengust 17 D. Stojanović
137 : Klop 12 Polšak 17 P. Stojanović 4 Centrih 33 Popović 20 Krljanović 3 Oparenović 14 A. Romih
146 : Klop 12 Čelofiga 3 Jeseničnik 9 Mešić 15 Renato 2 Tomažić Šeruga 8 Trifković 20 Amel Mujaković
153 : Notranje mere širina spredaj 1365 mm zadaj 1425 mm dolžina sedalnega dela prednji sedež 490 mm zadnja klop 460 mm premer volanskega obroča 375 mm posoda za gorivo 54 l.
156 : Klop 12 Vukalić 24 Manojlović 4 Novaković 14 Jašari 18 Rastoka 27 B. Stojanov
158 : Klop 12 Simčić 26 Komac 24 Lungu 10 Kršić 33 Birsa 15 Blaško
181 : Notranje mere dolžina od armaturne plošče do zadnjega naslonjala 165 cm širina spredaj 150 cm zadaj 146 višina nad sedežem spredaj 94 104 cm zadaj 94 cm vzdolžnica prednjega sedeža 93 115 cm zadnje klopi 84 60 dolžina sedala spredaj 52 cm zadaj 49 cm višina naslonjala spredaj 62 cm zadaj 62 cm
193 : 19.00 Iz občinskih klopi Seja Občinskega sveta Občine Tržič rep
194 : 17.00 Z OBČINSKIH KLOPI SEJA OS ŠKOFJA LOKA 10.11.05 1
211 : V našem primeru je bil kufer velik 320 litrov ob podrti klopi pa 1135 litrov
234 : Klop 22 Lirussi 29 L. Elsner 24 Arlić 26 Topić 17 Lucas 27 Marko Lunder 18 A. Hribar
235 : Klop 12 Simčić 8 N. Kovačević 21 M. Kovačević 11 Šabec 24 Lungu 15 Blaško 16 Jogan
```

```
67 : Klop 12 Fink 24 Brumen 21 Robnik 6 Duro 11 Urbanč 3 Pranjić 29 Travner
68 : Klop 29 Štelcer 21 Toplak 20 Drevenšek 26 Petek 23 Berko 9 Sladojević 14 Ljubanić
104 : Klop 27 Ivankov 88 Mihailov 3 Milanov 5 Vergilov 23 Džaeioba 18 M. Ivanov 19 Čilakov
127 : Klop 1 Strajanr 3 Grabus 4 Gerić 16 Majcen 23 Jelečević 29 Rujović 33 Grabić
134 : Klop 22 Rozman 24 Zadnikar 2 Skubic 8 Gerić 16 Majcen 32 B. Božić 20 Ntame
135 : Klop 1 Botonjić 3 Šporar 30 Mitraković 20 Leskovar 13 Rudonja 9 Ibraimi 32 Tiganj
136 : Klop 1 Vidmar 16 Mohar 3 Topić 14 Krcić 9 Šturm 27 Vengust 17 D. Stojanović
137 : Klop 12 Polšak 17 P. Stojanović 4 Centrih 33 Popović 20 Krljanović 3 Oparenović 14 A. Romih
146 : Klop 12 Čelofiga 3 Jeseničnik 9 Mešić 15 Renato 2 Tomažić Šeruga 8 Trifković 20 Amel Mujaković
156 : Klop 12 Vukalić 24 Manojlović 4 Novaković 14 Jašari 18 Rastoka 27 B. Stojanov
158 : Klop 12 Simčić 26 Komac 24 Lungu 10 Kršić 33 Birsa 15 Blaško
234 : Klop 22 Lirussi 29 L. Elsner 24 Arlić 26 Topić 17 Lucas 27 Marko Lunder 18 A. Hribar
235 : Klop 12 Simčić 8 N. Kovačević 21 M. Kovačević 11 Šabec 24 Lungu 15 Blaško 16 Jogan
```

Figure 2. The difference between agglomerative (left) and spectral clustering (right) Comparison of both clustering techniques on the word klop. They both returned similar clusters but spectral is more strict and contains only almost identical sentences.

Embedding	Cluster 1	Cluster 2	Cluster 3
word2vec	Po eni naj bi bil umrl za **rakom** na želodcu.	Menda se hrani z ribami **raki** in lignji sicer pa je o njenih navadah bolj malo znanega.	**Rak** od 22. junija do 22. julija Novica ki ste jo dolgo čakali vam bo polepšala dan.
word2vec	Tudi v Sloveniji se ravnajo po sprejetih normativih Evropskega združenja za boj proti **raku** dojk	Očiščene **rake** damo v omako dodamo timijan in na blagem ognju kuhamo 4 minute	**Rak** od 22. junija do 22. julija Lepo vam bo medtem ko se boste družili in pogovarjali z domačimi
word2vec	Informacije & podpora v boju proti črevesnemu **raku**	To velja še zlasti v odnosih z vodnimi **raki** škorpijoni ali ribami	**RAK** od 22. rožnika do 22. malega srpana
fastText	Namesto zanesljivega zaključka je David Špiler napravil prekršek v napadu najboljši mož Gorenja Vid Kavtičnik pa je s svojim 13. **golom** poskrbel za izenačenje na 28:28	Marina Park Danijele Savič je izgubila na gostovanju pri Vicar Goyi z 31:23 Savičeva pa je za svojo ekipo dosegla tri **gole**	-
fastText	Vrata so se odprla in stopila sta v veliko **golo** predsobo v kateri je na dolgih klopeh sedela množica ljudi	V bundesligi je Wilhelmshavener Boštjana Hribarja doma z 31:30 izgubil proti Nordhornu Hribar pa je za svojo ekipo dosegel štiri **gole**	-
fastText	Makedonija in Inter sta dolgo časa lomila odpor Maltežanov ki so drugi **gol** prejeli ko so se v želji po izenačenju povsem odprli	-	-

Table 3. Clusters of **Agglomerative clustering** using word2vec and fastText embeddings for word **rak** and **gol**. With fastText embeddings agglomerative clustering retured two clusters, where the second has only two sentences.

References

[1] Nataša Logar, Tomaž Erjavec, Simon Krek, Miha Grčar, and Peter Holozan. Written corpus ccGigafida 1.0, 2013. Slovenian language resource repository CLARIN.SI.

[2] Nikola Ljubešić and Kaja Dobrovoljc. What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy, August 2019. Association for Computational Linguistics.

[3] Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[4] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.