FRI Natural language processing course 2021/2022
Project report

University *of Ljubljana*
Faculty *of Computer and Information Science*

Cross-lingual sense disambiguation

Matej Miočić, Marko Ivanovski, Matej Kalc

Abstract

Classical 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. Contextual embeddings might be better off here, but how good are they in differentiating homonyms in their different meanings? We will try to tackle this problem for the Slovene language by first constructing our own corpora of sentences and then performing clustering or classification on them.

Keywords

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

Advisors: doc. dr. 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 the following, we will explain how we were able to form our corpora and then used non-contextual and contextual neural embeddings to try and solve this task.

Methods

We divided this task into two parts:

1. Corpus preparation
2. Clustering, 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 (**testing set**), which we thought could yield good results. In Figure 2 we show how many sentences we extracted for the selected words.

Clustering and analyses

After obtaining the corpus we started 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 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 vec-

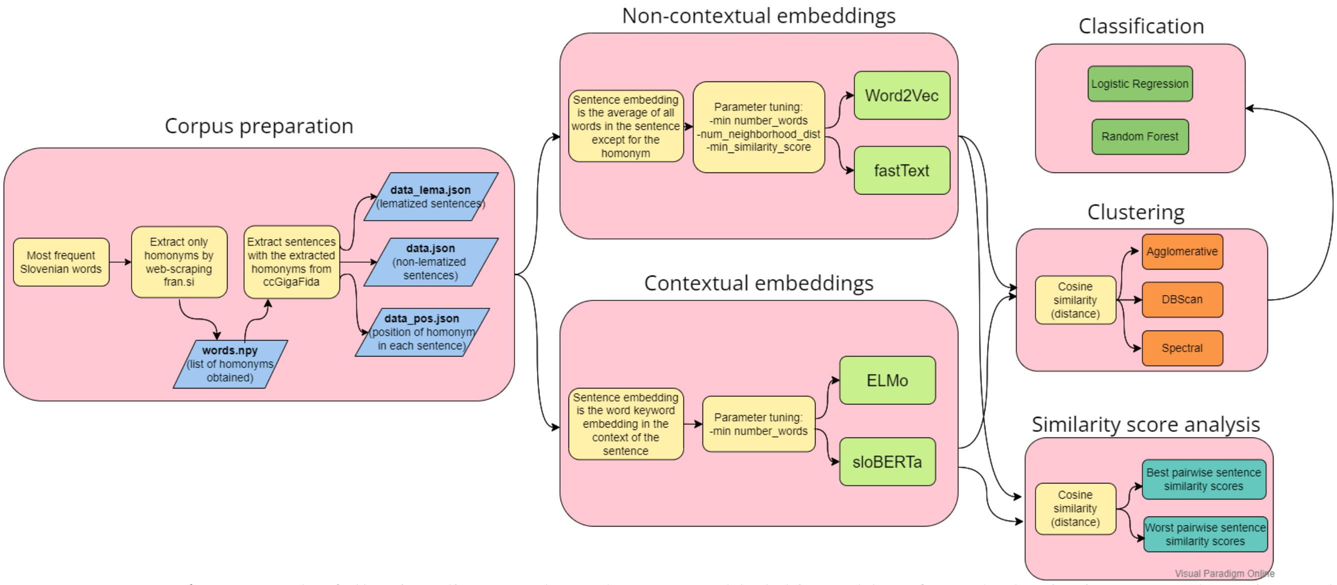

Figure 1. Flow diagram. The following diagram shows how we tackled this problem from the beginning up to clustering and classification.

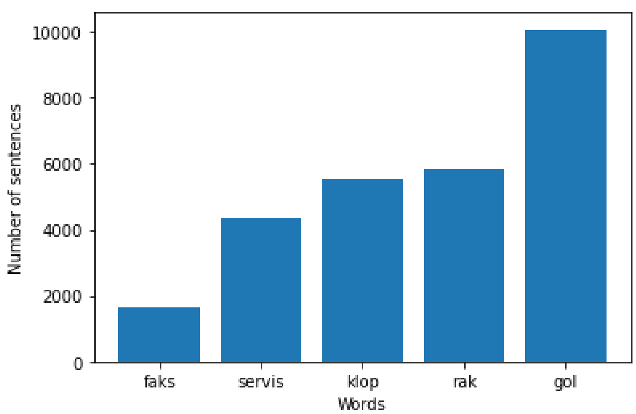

Figure 2. Number of sentences extracted for chosen words.

tors. 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]. 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.

Next, we also tried contextual embeddings pre-trained in the Slovene language like ELMo [5] and sloBERTa 2.0 [6]. Obtaining the sentence embeddings was more straightforward here since we only pass the sentence to the model and obtain the embedding for our homonym keyword. In this approach, we only had one parameter: the minimum number of words that a sentence must have.

After obtaining the embedding matrix for all sentences (for both approaches) 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 that had the best and worst similarity scores.
- We also clustered the embeddings for a single word.

Classification

Our final goal was to classify new sentences with the correct meaning of the keyword. With obtained clusters for each meaning of the word, we used the embeddings to train a binary classifier. We annotated the first cluster with label 0 and the second cluster with label 1. For each sentence embedding, we created a target variable with the cluster label.

This assumes that clusters are correct for each embedding, but sadly that is not the case. Therefore we additionally annotated 92 sentences for word **gol** to see how the models perform on a correctly labeled test set. Results are shown in the next section (see Figure 3 and Table 1).

Results

Non-contextual embeddings

In the following analysis, for the non-contextual embeddings, 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 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 4 and 3.

Agglomerative Clustering

Word2vec and fastText embeddings were clustered using linkages: single, average, and complete. In the case of ”rak”, word2vec embedding clusters are divided well into 3 groups, where each represents a certain meaning, which is the desired result (Table 2). For the same settings, fastText embedding did not work that good. 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.

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 4). 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 was 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 [7]. 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.

Contextual embeddings

For the contextual embeddings, we used **8** as the minimum number of words required in a sentence. There was no need to use the `minimum similarity required` parameter since both models worked better from the start. Finally, since we noticed that hierarchical clustering worked better in most cases we only focus on those results.

ELMo

ELMo initially seemed to work well. Testing it initially on the words **gol** and **klop** gave surprisingly almost perfect clusters (Table 5). For the word **rak** it could not separate the 3 clusters well but it could separate the disease from the animal, although not perfectly. The clusters for **golf** and **faks** were not the cleanest although we saw that at least could differentiate between the two meanings. Regarding worst and best score analysis the same observation as before was made. Best scores are sentences that contain the word in the same meaning but this did not always hold for worst similarity scores(Table 6).

sloBERTa 2.0

For word embeddings with sloBERTa, we summed the last 4 layers for each token.(We also experimented by using second to last layer but the results were similar.) When using sloBERTa all clusters were not as good, because the majority of the sentences are in one cluster. The majority of times the second cluster contains all the sentences that start with the homonym. An example of such clusters is shown in Table 5. Although we expected more, it seems for now ELMo gives the best embeddings.

Classification

To obtain results we performed a 5-fold cross validation repeated 10 times for each embedding model. We used logistic regression and random forest. In Figure 3 we show means of log loss and their 95% confidence intervals for each word and each embedding model. Since every embedding model produced different clusters and therefore labeled sentences differently, we also show the baseline for every embedding model. We can see that the classifiers outperformed the baseline every time. We observe almost 0 log loss when classifying with sloBERTa since the clusters can be easily classified but that does not mean they are correctly labeled. We observed that one cluster contained sentences that began with the keyword no matter the meaning. This is why we decided to correctly annotate 92 sentences for word **gol**. With this correctly labeled validation set, we can see if any of the models that were trained on clusters can outperform the baseline.

Table 1. Classification accuracy with standard deviation for all four embedding models on 92 annotated sentences for word **gol**.

	Logistic regression	Random forest
Word2Vec	0.86 ± 0.00	0.86 ± 0.00
fastText	0.86 ± 0.00	0.86 ± 0.00
ELMo	$\mathbf{0.92} \pm 0.01$	0.90 ± 0.01
sloBERTa	0.86 ± 0.00	0.86 ± 0.00
baseline	0.86	

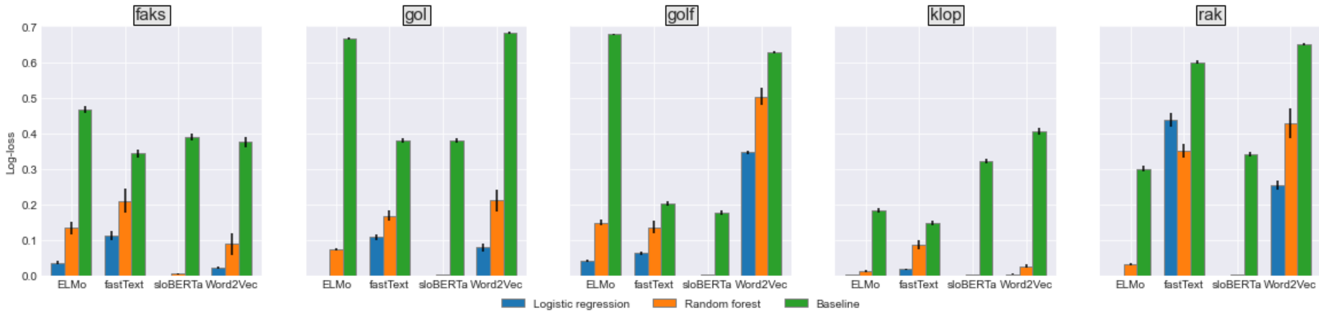

Figure 3. Bar plots of log loss with 95% confidence intervals for predicting sentences with logistic regression and random forest with 5-fold cross-validation repeated 10 times of 5 words: **faks, gol, golf, klop** and **rak** with clusters obtained with 4 used embedding models: **ELMo, fastText, sloBERTa** and **Word2Vec**.

Just like before we performed a 5-fold cross-validation repeated 10 times. In Table 1 we observe that only clusters with ELMo embeddings managed to get a higher classification accuracy than the baseline. That means that the other clusters were not successfully created and trained in the model to predict only the majority.

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 in the same context clusters. We also saw that clustering results depend highly on the pre-trained model. ELMo outperformed all the others. The ELMo contextual embeddings always resulted in at least one pure cluster if not even two on our test homonyms. One difficulty we noticed in clustering and classification is having a very imbalanced sentence set (with different contexts) for one keyword. We have shown classification results on clusters produced with different embedding models. Since some clusters were unsuccessful, we annotated 92 sentences to truly see how accurate the classifications are. We have shown that only ELMo managed to beat the baseline for these 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] Matej Ulčar. ELMo embeddings models for seven languages, 2019. Slovenian language resource repository CLARIN.SI.

[6] Matej Ulčar and Marko Robnik-Šikonja. Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0, 2021. Slovenian language resource repository CLARIN.SI.

[7] 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.

Appendix

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

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

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

Embedding	Cluster 1	Cluster 2
ELMo	Na zatožni **klopi** se mora braniti zaradi spolne zlorabe mladoletnika.	Pinceta za **klope** je iz nerjavečega jekla s posebej brušenimi prijemalkami.
ELMo	Zadnja **klop** pri audiju a3 seveda ni tako udobna in prostorna kot v osmici.	**Klopa** ne zmečkajte temveč ga odvrzite v koš za gospodinjske.
ELMo	Bilo je 2:1 dina šeremeta tokrat ni bilo na gostujoči **klopi** za rezervne igralce.	Hitra odstranitev vsesanega **klopa** bistveno zmanjša možnost okužbe.
sloBERTa	Foto4 Miš ali srna je najpogostejši gostitelj **klopov** človek pa naključni	**Klop** Troha Burgar Nišandžič Hodžič Krcič Novak Ovijač
sloBERTa	Srečko Lampe pika **klopa** ni prijavil niti ni bil na pregledu v ambulanti Učnega centra Vrhnika	**Klop** Handanovič Ilič Pečnik Sešlar Šiljak Cimerotič Brečko
sloBERTa	In slabo uslugo bi si naredili če bi se začeli v vladnih **klopeh** ozirati po domačih izdajalcih ki so odtegnili svoj glas	**Klop** 29 Štelcer 30 J. Emeršič 13 Prejac 27 Chietti 17 Bosilj

Table 5. Clusters of **Agglomerative clustering** using ELMo and sloBERTa embeddings for word **klop**.

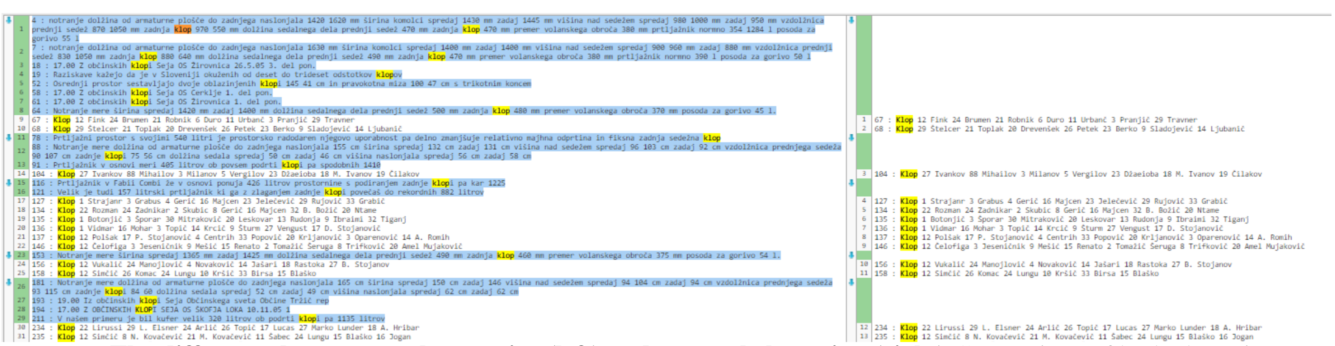

Figure 4. The difference between agglomerative (left) and spectral clustering (right) Comparison of both clustering techniques on the word klop. They both returned similar clusters but spectral is more strict and contains only almost identical sentences.

Score	Sentence 1	Sentence 2
0.985	**Golf** b letnik 1987 registriran do 12 98 rdeč prodam 068 74 020.	**Golf** d letnik 1987 registriran do 28.3.1999 prodam 068 325 334.
0.985	Volkswagen **golf** 1.9 tdi highline am 1 02.	Volkswagen **golf** 1.9 tdi dsg am 19 04.
0.984	**Golf** jx d letnik 1988 moder 3v zelo lepo ohranjen brezhiben prodam.	**Golf** jx d letnik 1988 registriran do 2 98 bel 3v dobro ohranjen prodam.
0.154	..nižjega ali srednjega cenovnega razreda malo je **golfov** ne prav veliko manjših renaultov nissanov toyot hyundaijev in podobnih vsakdanjih uporabnikov naših cest.	Celoten izkupiček dražbe je skupaj s sredstvi zbranimi na **golf** turnirju z akcijo nearest to the pin znašal 2,5 milijona tolarjev.
0.149	Predsednik prvega celjskega **golf** kluba je borut sedovnik.	Morda ste bili presenečeni da ste v prejšnji številki videli toliko **golfov** na kupu...
0.132	Enega od **golfov** iz omejene serije teh ljudskih športnikov smo preizkusili tudi mi.	Zaradi opisane ameriške logike je znanost prenehala biti posvečen azil elitni intelektualni **golf** klub...

Table 6. Three most and three least similar sentences using **ELMo** for word **golf**.

Score	Sentence 1	Sentence 2
0.999	**GOLF** JX D letnik 1990 registriran do 9 2000 5V bel dobro ohranjen prodam 068 73 105	**GOLF** JX D letnik 1988 registriran do 2 98 bel 3V dobro ohranjen prodam 068 73 105
0.999	**GOLF** PLUS 1,9 TDI 66 kW 18.320 EUR	**GOLF** PLUS 1,9 TDI 66 kW 4.487.000 SIT
0.999	**GOLF** Variant Basis 1.9 TDi 74kW 4MOT 3.991.890	**GOLF** Variant Basis 1.9 TDi 66kW Avt. 3.835.677
-0.023	**GOLF** letnik 1979 dobro ohranjen prodam za 144.000 SIT 041 652 056	... 20.00 DSF Novinarski center 20.15 Wimbledon 22.15 DSF Novinarski center 22.30 LaOla 23.00 Nogomet 0.50 **Golf** US Open 1.50 Best Direct
-0.022	**GOLF** letnik 1979 dobro ohranjen prodam za 144.000 SIT 041 652 056	Američanka je namreč v Londonu tokrat prvič v karieri predstavila svojega spremljevalca igralca **golfa** Henryja Augusta Kuehneja II ...
-0.027	**GOLF** TD 1.9 letnik 1992 registriran do 2 2000 kovinsko svetlo zelen 5V alu platišča športni sedeži daljinsko CZ 155.000 km prodam 068 322 407	... in Hercegovine B. M. nato pa še v **golfa** in beemweja ki sta ju vozila hrvaška državljana 21 letni S. V. ...

Table 7. Three most and three least similar sentences using **sloBERTa** for word **golf**.