POS, ANA and LEM:
Word Embeddings Built from Annotated Corpora Perform Better

Attila Novák, Borbála Novák

Pázmány Péter Catholic University Faculty of Information Technology and Bionics
MTA-PPKE Hungarian Language Technology Research Group
Budapest, Práter u. 50/a.
{novak.attila, novak.borbala}@itk.ppke.hu

Abstract. Word embedding models have been popular and quite efficient tools for representing lexical semantics in different languages. Nevertheless, there is no standard for the direct evaluation of such models. Moreover, the applicability of word embedding models is still a research question for less resourced and morphologically complex languages. In this paper, we present and evaluate different corpus preprocessing methods that make the creation of high-quality word embedding models for Hungarian (and other morphologically complex languages) possible. We use a crowd-sourcing-based intrinsic evaluation scenario, and a detailed comparison of our models is presented. The results show that models built from analyzed corpora are of better quality than raw models.
Keywords: word embeddings, intrinsic evaluation, morphologically rich languages

1   Introduction and Related Work

Finding a good representation of words and lexemes is a crucial task in the field of NLP. Neural word embeddings (WE) have proved to be efficient for such tasks [1,2,3].

Various evaluation methods have been proposed for evaluating such models. Some studies use extrinsic evaluation by measuring the effect of using WE's as features in specific tasks. Extrinsic evaluation is important, however, it does not reflect the quality of the embedding model itself, only its effect, which depends on the nature of the downstream task it is utilized in. In most cases, WE models are evaluated in word similarity tasks [4,3] measuring the correlation of the similarity score various embedding models assign to pairs of words to scores assigned by humans (stored in resources called query inventories). However, as Schnabel et al. [5] point out, in this case the problem is that scores from different rankings with different ranges cannot in general be compared, and aggregating such scores may lead to false results. In order to be able to use rank correlation as a metric, a gold similarity of all word pairs in the set including pairs of

completely unrelated words would be needed, however, data for only a subset of such pairs is included in these resources. Thus, in order to evaluate and compare WE models in general, they suggest the use of a balanced test set with respect to word frequency, abstractness and part-of-speech.

Most studies focus on the application of embedding models to English or other languages with simple morphology, where the moderate number of different word forms and the relatively fixed word order fit well the theory behind these models. Query inventories are also in general available only for a few languages, mainly English. The dimension along which evaluation is generally performed is the comparison of different implementations of WE models and parameter settings used for training the models on the same corpus. However, in the case of morphologically complex languages, the question of how to handle different word forms of the same lemma has to be considered as well. Our research question is motivated by the fact that, in general, statistical language models built for morphologically complex languages suffer from data sparseness problems: e.g. word-based language models for a morphologically complex language have a notoriously higher perplexity than those for English when created from a corpus of the same size [6].

Ebert and his colleagues experiment with creating lemmatized and stemmed models for several morphologically complex languages (including Hungarian) [7] and conclude that the lemmatized models perform best in word similarity tasks. They introduce a WordNet-based mean reciprocal rank (MRR) metric based on the position of the first item on the nearest neighbor (NN) list retrieved from the embedding model that can be reached from the query word within two steps along some relations present in WordNet. However, the quality and the low and biased lexical coverage of the Hungarian WordNet (HuWN) [8] does not make it a reliable gold standard of semantic relatedness of words. A high portion of HuWN consists of proper names (83% of the 19400 noun synsets, 38% of all synsets), while the coverage of frequent common words is not very good. Another problem is that HuWN follows the structure of the English (Princeton) WordNet (v2.0) containing many "ghost" synsets that do not have any corresponding word in Hungarian (such nodes make up 6.7% of all synsets). These problems may partly account for the fact that the mean reciprocal rank scores obtained in [7] for Hungarian were so much lower than those obtained for the other languages involved in that experiment. We included the HuWN in our experiments described in Section 3 in order to demonstrate that it is not a very good benchmark to compare the performance of WE models against.

The goal of this paper is to investigate the performance of WE models applied to a morphologically complex agglutinative language with free word order, Hungarian, and to perform an exhaustive intrinsic evaluation comparing the effect of different preprocessing scenarios applied to the training set used when building the models. We crucially rely on human evaluation here, to be performed by native speakers in order to get reliable results. This made us limit our investigation to Hungarian.

2 Embedding Models for Hungarian

kenyerek(2270) 'breads' SURF	Vakkalit(5) 'Vakkali.Acc' SURF
kiflik(349) 'bagels'	tevedesnek(5) 'as a mistake'
zsemlék(283) 'buns'	áfa-jának(7) 'of its VAT'
lepények(202) 'pies'	mot-nak(5) 'mot.Dat'
pogácsák(539) 'scones'	Villanysze(5) 'Electrici(an)'
pékáruk(771) 'bakery products'	oktávtól(5) 'from octave'
péksütemények(997) 'pastry.pl'	Isten-imádat(5) 'worship of God'
sonkák(613) 'hams'	Nagycsajszi(5) 'Big Chick'
tészták(2466) 'pasta.pl'	-fontosnak(7) '-as important'
kalácsok(277) 'cakes'	tárgykörbôl(5) 'frôm the subject'
kenyér(147000) 'bread' ANA	Vakkali(23) ANA
hús(136814) 'meat'	Ánanda(321)
kalács(10658) 'milk loaf'	Avalokitésvara(39)
rizs(31678) 'rice'	Dordzse(270)
zsemle(6690) 'roll'	Babaji(82)
pogácsa(11066) 'bisquit'	Bodhidharma(210)
sajt(46660) 'cheese'	Gautama(574)
kifli(9715) 'croissant'	Mahakásjapa(25)
krumpli(37271) 'potato '	Maitreya(426)
búzakenyér(306) 'wheat bread'	Bódhidharma(115)

Table 1: Top NN's of a frequent and a rare word from the SURF and the ANA model. Numbers are corpus frequency.

We built four types of models using the word2vec tool using the CBOW model. As a training corpus, we used a 1.2-billion-word raw web-crawled corpus of Hungarian [9]. When building each model, context window was set to 10 words, dimensions to 300, and minimal word occurrence limit to 5. Then we applied different types of preprocessing to the corpus in order to mitigate data sparseness effects due to agglutination. The same strategies can be applied to any other morphologically rich language having a morphological analyzer/tagger/lemmatizer available.

First, we built a model from the tokenized but otherwise raw corpus (SURF). This model is able to represent morphological analogies. E.g. the similarities of the word pairs *jó – rossz* 'good – bad' and *jobb – rosszabb* 'better – worse' are much higher in this model than if we compare the suffixed form and its lemma, i.e. *jó – jobb* 'good – better', and *rossz – rosszabb* 'bad – worse'. The first two examples in Table 1 show the nearest neighbors retrieved for some surface word forms. The model represents both semantic and morphosyntactic similarities. The top-n list for *kenyerek* 'bread.plur' has mostly pastries in plural. However, since due to agglutination, there is a high number of possible surface forms of the same lemma (there are 197 different inflected forms for the lemma

kenyér 'bread' in the corpus) the model is sensitive to data sparseness effects, since the contexts a word is used in are divided between the different surface forms of the same lemma. The `SURF` model is often not capable to capture the semantics of rare word forms reliably (e.g. the most similar entries for *Vakkalit* 'Vakkali.Acc' are completely unrelated forms in Table 1, col. 2).

Using a morphologically annotated version of the corpus, **we also created a lemmatized model (`LEM`)** in order to bias our model towards representing semantic rather than morphosyntactic and syntactic relatedness and to mitigate data sparseness issues. The annotation was created using the PurePos part-of-speech tagger [10] which also performs lemmatization using morphological analyses generated by the Hungarian Humor morphological analyzer (MA) [11,12,13].

Since using lemmata only to build the embedding model may overemphasize semantic relations while suppressing syntactic distributional regularities in the data, **we also created another analyzed (`ANA`) model** following the method described in [14]. A similar method was applied when creating a morpheme-based MT system for Hungarian to solve morphology-related data sparseness problems in [15]. Here we used the same morphologically annotated corpus, but instead of keeping only the lemmata, each word form in the corpus was represented by two tokens: a lemma token followed by a morphosyntactic tag token. The following example shows the representation of the sentence *Szeretlek, kedvesem.* 'I love you, sweetheart.' preprocessed this way:

```
szeret #V.1Sg.>2Sg , #, kedves #N.Poss1Sg
love    V.[I, you] ,    dear    N.[my]
```

Since the tags are kept in the actual context of the word they belong to, the morphosyntactic information carried by the inflections still has a role in determining the embedding vectors. On the other hand, data sparseness is reduced, because the various inflected forms are represented by a single lemma. The second two columns of Table 1 show some examples of top-n lists generated by this model. In contrast to the `SURF` model, the `ANA` model is capable of capturing the semantics of rare lexical items because lemmatization alleviates data sparseness problems and morphosyntactic annotation provides additional grammatical information (compare columns 2 and 4). The most similar entries of *Vakkali* 'Vakkali' in the `ANA` model clearly indicate that the model managed to capture the fact that this is the name of a Buddhist personality.

One of the main drawbacks of raw WE models is that they are not able to handle homonymy, i.e. if a single word form has several distinct meanings, the model assigns a single vector to the word either biased towards one meaning (if that meaning is prevalent in the corpus) or representing a mixture of several meanings. In order to handle this problem at least in those cases, where the different meanings of a word form correspond to different parts of speech, **we also created a modified (`POS`) version** of the `ANA` model keeping the main PoS tag of each word attached to the lemma and only the rest of the morphosyntactic tag was detached. This model assigns a different representation to homonyms having different PoS. The example sentence, *Szeretlek, kedvesem.* 'I love you, sweetheart.' looks like the following in the PoS version of the training corpus:

```
szeret#V #1Sg.>2Sg ,#, kedves#N #Poss1Sg
 love.V    [I, you]  ,   dear.N     [my]
```

3   Experiments

Since no query inventory containing human-assigned word relatedness scores for
Hungarian exists that could be used to perform intrinsic evaluation of Hungar-
ian WE models, we decided to follow the method described in [5] and perform
comparative intrinsic evaluation of the models setting up a crowdsourcing web
page to solicit human evaluation of the different WE models. We asked partic-
ipants to rank the systems according to relatedness of the words returned by
the systems to the query word. In addition to the `ANA`, `POS` and `LEM` models, a
modified version of the `SURF` model, an off-the-shelf skip-gram-based model built
from a raw corpus (`SGL`) and one based on HuWN were included.

Models built from raw corpora (such as the `SURF` model) contain suffixed
word forms, and these appear on the NN lists output by these models. In order
to be able to compare the output of these models to those built from prepro-
cessed versions of the corpus, the output of raw models was postprocessed. When
nearest neighbors from the original `SURF` model are retrieved for a query word,
the result list contains various inflected word of the lemma of the query word or
other related words in the top-n list. In order to get the top-n nearest lexemes
for the query word instead (and thus make the list comparable to the output of
the other models), the resulting list was lemmatized using the MA and repeated
occurrences of the same lemma were filtered out, resulting in k nearest lemmata
(`SURFL`).

Our models were built from a 1.2-billion-word Hungarian web corpus. There
is a freely available model built from a larger, 4.6-billion-token, raw Hungar-
ian webcorpus [16]. While we used the CBOW algorithm to train our models,
that model was created using the skip-gram word2vec model, with the default
parameter settings (200 dimensions, window size=5). We considered this as a
baseline model. We included the lemmatized and filtered output of this model
(`SGL`) in our evaluation in order to compare our results to an off-the-shelf model
trained using a different algorithm and lower dimensions and narrower window
on a larger but different training corpus.

We also included a model derived from HuWN (`WN`). Following the method
described in [7], we listed words in the HuWN for each query word that could
be reached by an at most 3-steps-long path along any WN relation. The list was
sorted by distance from the query word with synonyms of the query word being
at distance zero.

3.1   The Set of Queries

A set of query words for English balanced with respect to word frequency, ab-
stractness and part-of-speech were published in [5][1]. This resource included 100

---

[1] The resource is available online at `http://www.cs.cornell.edu/~schnabts/eval/`

query words. We translated these words (selecting a translation matching the sense supertag and the part of speech given in the original query inventory) and checked the corpus frequency range of the translation. If the translation was not in the same range, then we changed it to another word with the same part-of-speech and semantic category from the proper frequency range. We added 7 words exhibiting homonymy crossing a part of speech boundary, like *vár* 'castle NN' vs. 'wait/expect VB' or *reggeli* 'breakfast NN' vs. 'morning JJ'.

3.2 The Ranking Task

We asked human annotators to rank the six models (`ANA, SURFL, POS, LEM, SGL, WN`) through a dedicated web interface. In each turn for each annotator, a query word was shown from our list and either 1, 5 or 9 words from each model. The order of the models was randomized in each test case. The result lists were extracted from the k-nearest neighbor lists generated for the query word by each model. The 1, 5 or 9-word-long sequences started either at rank 1 or at rank 30 (in each turn, words from the same range were displayed for each model).[2] There was one exception: in cases when the starting position was 30 for the other models, the `WN` model was left out of the comparison, because the lists generated from HuWN were shorter than 30 items for 97 of the 107 selected words. The coverage of HuWN was rather low anyway: only 70 of the 107 words were covered (65%). If any of the models did not have any result for the given query, an empty list was displayed.

Table 2 shows an example question for the query word *nyúl (főnév)* 'rabbit (noun)' presenting a one-word-long list from each model to the annotators[3].

nyúl (főnév) 'rabbit (noun)'					
macska	szalad	nyúlkál	nyúlféle	malac	dörgölőzik
'cat'	'run'	'touch'	'lagomorph'	'pig'	'rub'

Table 2: An example question asking the annotators to rank the presented words according to their similarity to the query word *nyúl (főnév)* 'rabbit (noun)'

The annotators had to rank the lists from worst to best according to their intuition of relatedness. For ranking, they could assign a number from 1 to 99 (higher score better) to each system output (0 was given for empty lists by default) and ties were allowed. Annotators were allowed to pass to the next query without ranking when they did not know the word or found ranking the lists too difficult. When evaluating the rankings, absolute values of the scores

---

[2] Although testing at NN rank 30 may seem odd at first, word2vec output even at around rank 2000 often perfectly makes sense. Most entries at around rank 2000 for *macska* 'cat' or *nyúl* 'rabbit' are animals.

[3] *nyúl* is also a verb in Hungarian 'reach for'. The verbal sense in 4 times more frequent in the training corpus, dominating the vector representation for most models.

were ignored. Annotators were also informed in the instructions about the fact that the scores are only used for ranking the lists.

The metric we used for ranking the models was win ratio: the average number of times that the output of a model was judged to be better than another model for the same input. This means the pairwise comparison of the output generated by each model for each query word based on annotator ranking. Each winner scores a point. The final score is wins/comparisons. This method has been used to rank the output of machine translation systems in human evaluation since the beginning of the WMT workshop series, and it has been found to be the most reliable method of comparing the quality of widely differing systems [17]. Rankings were calculated for different values of different parameters, such as the query word frequency range, part of speech, and semantic category (abstractness) of the query word, and the position in the nearest neighbor list (NN 1 or NN 30) where the lists were taken from. The `WN` model was left out of comparison at NN position 30.

Even though the annotators were all native speakers of Hungarian, we introduced a test phase into the system. I.e. each new annotator was given 5 test words randomly placed among the first 10 questions. These test cases contained easy-to-order words and we compared the answers to a gold standard ranking. If an annotator did not manage to reliably distinguish at least distributionally close words and unrelated ones in at least 80% of the test cases, then his or her results were not included in the evaluation. In the end, 15 different annotators provided useful answers each of them submitting 14 to 102 different rankings. We got at least 3 rankings for each query word both at NN position 1 and 30.

4   Results

Results of the evaluation are shown in Figure 1. One obvious fact that can be seen in the graphs is that the off-the-shelf 200-dimension skip-gram (`SGL`) model [16] performed worst for each condition (p value < 0.05) except for adjectives and very rare words (frequency ≤ 500) where `WN` performed just as badly due to lack of coverage, and the PoS ambiguity *(posamb)* condition (see Figure 1d and its discussion below). NN lists from the `SG` model (i.e. the `SGL` model before lemmatization) contain many "intruders" also for frequent words (e.g. *búzalisztből* 'from wheat flour', *sütéséhez* 'for its baking', *karfiolból* 'from cauliflower' etc. among the top 20 nearest neighbors of *kenyerek* 'loaves of bread'), while the output for rare words often seems to be just all junk. Since the `SG` model was not created by us, we do not know what makes this model perform so poorly (despite the fact that it was created from a significantly larger corpus). It is not probable that the difference is due to the different algorithm (skip-gram vs. CBOW).

Absolute ranking of the models is shown by the *all frq* condition in Figure 1a. All models we built performed significantly better than HuWN, which confirmed our doubt about using HuWN as a benchmark for the evaluation of unsupervised

(a) frequency

(b) position
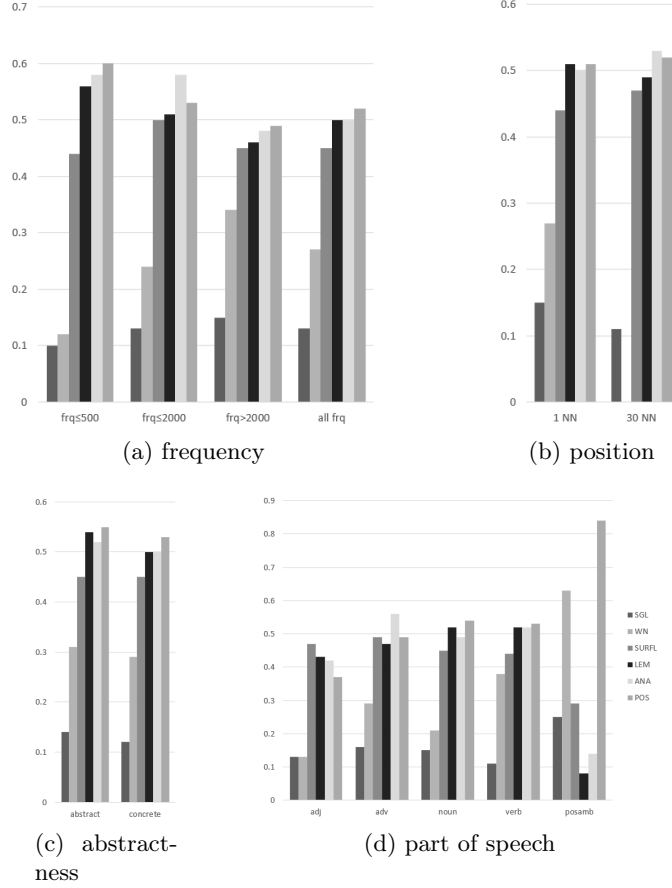
(c) abstractness

(d) part of speech

Fig. 1: Performance (win ratio) of the models for different conditions.

embedding resources for Hungarian, its performance being especially dismal for rare words due to the lack of coverage.

As it can be seen in Figure 1a, building the WE models from an annotated/lemmatized corpus did improve performance for infrequent words (with word form frequency below 500). For frequent words, on the other hand, the performance gain was less pronounced.

When taking the position in the nearest neighbor lists into account (Figure 1b), the already low quality of the skip-gram model seems to further deteriorate at position 30.

A more interesting result is that the quality of the output of the `ANA` and `POS` models seems to remain higher at NN 30 than that of the `LEM` model. While the lemmatized model seems to handle data sparseness well (see the quality gain at low frequencies), it seems to overrepresent semantic relatedness at the

expense of representing grammatical similarity. The NN lists coming from the `LEM` model are often quite heterogeneous concerning part of speech. The `ANA` model represents grammatical (syntactic) similarity much better because we kept some representation of the morphosyntactic features in the context when training the embedding vectors for the lemmas. Furthermore, the `POS` model can even distinguish different senses of homonymous words, as far as ambiguity can be resolved at the level of part of speech tagging. These models seem to be a viable compromise between the word-form-based models built from raw corpora, which, while they represent morphosyntactic knowledge, suffer from data sparseness problems for rare words, and the wildly semantic lemmatized models, which fail to represent much of the (morpho)syntactic information in the data. This is so despite the fact that using the same window size parameter corresponds to only half the context when building the `ANA` model compared to that used when building the `LEM` and `SURF` models[4].

As for abstractness, the relative performance of the models turned out to be almost independent of this dimension (Figure 1c). The results for different parts of speech can be seen in Figure 1d. One clear trend is that the models trained on an analyzed/lemmatized corpus clearly performed better for verbs and nouns. This is not surprising, given that these are the most massively inflected categories in Hungarian. The same did not turn out to be true for adjectives: here, the surface model performed better. Note that while adjectives can take the same inflections as nouns in Hungarian in addition to comparatives and superlatives, suffixed forms for adjectives are rare in Hungarian corpora due to the lack of marked agreement within the noun phrase.

The `POS` model performed at its best when tested on homonymous words one sense of which has a different part of speech than the other (see the *posamb* condition in Figure 1d). The second best performer was the `WN` model in this task, as HuWN also contains part-of-speech information. The `POS` model performed better because it has better coverage. This was the only test where the `LEM` and `ANA` models performed worse than both raw-corpus-based models (`SGL` and `SURFL`).

5 Conclusion

In this paper, we presented the results of the crowd-sourcing-based intrinsic evaluation of WE models created for Hungarian, a morphologically complex agglutinative language. We built four models with different preprocessing of the same corpus using the CBOW word2vec model. We included two additional

---

[4] Each word is represented by exactly two tokens in the `ANA` model, thus the same context window only covers half as many words. The same applies to inflected word forms in the `POS` model, while noninflected words are represented by only a single token in that model. Note that this effect is mitigated by the fact that word2vec downsamples frequent word forms (among them frequent tags) when creating the model. This corresponds to an effective window size expansion.

models in the evaluation: an off-the-shelf skip-gram model built on a larger raw tokenized corpus and a model based on HuWN.

We found the models utilizing morphological annotations perform better than raw surface-from-based models mitigating data sparseness problems following from agglutination. The novel `ANA` and `POS` models were found to perform even better than the lemma-based model, because, due to the fact that morphosyntactic information is preserved in the context when the models are trained, these models better represent grammatical similarities in addition to semantic ones. Although these models are simple (like Mikolov's CBOW model itself), they perform surprisingly well. The `POS` model is capable of capturing sense distinctions that correspond to PoS distinctions. Our experiments also showed that the quality and especially the coverage of the HuWN resource is not good enough to use it as a benchmark for evaluating Hungarian WE models, as the models themselves were found to perform significantly better under all conditions tested.

Acknowledgments

References

1. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
2. Mikolov, T., Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA. (2013) 746–751
3. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, Maryland, Association for Computational Linguistics (2014) 238–247
4. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States. (2013) 3111–3119
5. Schnabel, T., Labutov, I., Mimno, D.M., Joachims, T.: Evaluation methods for unsupervised word embeddings. In Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y., eds.: EMNLP, The Association for Computational Linguistics (2015) 298–307
6. Popel, M., Mareček, D. In: Perplexity of n-Gram and Dependency Language Models. Springer Berlin Heidelberg, Berlin, Heidelberg (2010) 173–180

7. Ebert, S., Müller, T., Schütze, H.: Lamb: A good shepherd of morphologically rich languages. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, USA (2016)

8. Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., Váradi, T.: Methods and results of the Hungarian WordNet project. In: Proceedings of The Fourth Global WordNet Conference. (2008) 311–321

9. Endrédy, I., Prószéky, G.: A pázmány korpusz [the 'pázmány' corpus]. Nyelvtudományi Közlemények **112** (2016) 191–206

10. Orosz, G., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013), Hissar, Bulgaria, INCOMA Ltd. Shoumen, BULGARIA (2013) 539–545

11. Novák, A.: Milyen a jó Humor? [What is good Humor like?]. In: I. Magyar Számítógépes Nyelvészeti Konferencia [First Hungarian conference on computational linguistics], Szeged, SZTE (2003) 138–144

12. Prószéky, G., Kis, B.: A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. ACL '99, Stroudsburg, PA, USA, Association for Computational Linguistics (1999) 261–268

13. Novák, A.: A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014) 1068–1073 ACL Anthology Identifier: L14-1207.

14. Siklósi, B.: Using embedding models for lexical categorization in morphologically rich languages. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, Springer International Publishing, Cham. (2016)

15. Laki, L., Novak, A., Siklósi, B.: English to hungarian morpheme-based statistical machine translation system with reordering rules. In: Proceedings of the Second Workshop on Hybrid Approaches to Translation, Sofia, Bulgaria, Association for Computational Linguistics (2013) 42–50

16. Szántó, Z., Vincze, V., Farkas, R.: Magyar nyelvu szó- és karakterszintu szóbeágyazások. In Tanács, A., Varga, V., Vincze, V., eds.: XIII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport (2017) 323–328

17. Bojar, O., Ercegovčević, M., Popel, M., Zaidan, O.F.: A grain of salt for the wmt manual evaluation. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. WMT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 1–11