Using embedding models for lexical categorization in morphologically rich languages

Borbála Siklósi

Pázmány Péter Catholic University,
Faculty of Information Technology and Bionics,
50/a Práter street, 1083 Budapest, Hungary
siklosi.borbala@itk.ppke.hu

Abstract. Neural-network-based semantic embedding models are relatively new but popular tools in the field of natural language processing. It has been shown that continuous embedding vectors assigned to words provide an adequate representation of their meaning in the case of English. However, morphologically rich languages have not yet been the subject of experiments with these embedding models. In this paper, we investigate the performance of embedding models for Hungarian, trained on corpora with different levels of preprocessing. The models are evaluated on various lexical categorization tasks. They are used for enriching the lexical database of a morphological analyzer with semantic features automatically extracted from the corpora.

1 Introduction

Finding a good representation of words and lexemes is a crucial task in the field of natural language processing. The question is what type of representation to use that is able to model the distributional patterns of words including their meaning and their morphosyntactic and syntactic behavior. For English, the use of continuous vector space representations have recently replaced the manual creation of such resources as well as that of sparse discrete representations learned from analyzed or raw texts. The neural-network-based implementations of these continuous representations have proved to be efficient as shown in several publications [7, 9, 2]. Most studies, however, focus on the application of these models to English, where the moderate number of different word forms and the relatively fixed word order fit well the theory behind these models. The goal of this paper is to investigate the performance of word embedding models applied to Hungarian, an agglutinating language with free word order.

The motivation of our investigation is, however, twofold. First, our goal was to explore the semantic sensibility of embedding methods for a language more complex than English, i.e. whether it is able to locate words consistently in the semantic space when trained on Hungarian texts. On the other hand, the possibility of using the results to augment the stem database of a Hungarian morphological analyzer with semantic features was also investigated [13]. Semantic features may affect the morphological, syntactic and orthographic behavior of words in

Hungarian with certain constructions being applicable only to specific classes such as colors, materials, nationalities, languages, occupations, first names etc. and words falling into these categories could be identified and collected from the corpus resulting in exhaustive lists, which could not have been built manually. Moreover, lexical semantic categories extracted by the models can be used to enrich the annotation of argument slots in verbal subcategorization frames like the ones in [5] with further semantic constraints.

The structure of this paper is as follows. First, some of the main characteristics of Hungarian is described, which demonstrates the complexity of the given task. Then a brief summary of related work and continuous embedding models is presented. In the following sections our experiments are described regarding building different models from differently preprocessed corpora, and their use in the task of extracting semantic categories. Finally, both qualitative and quantitative evaluations of the results are presented along with describing some methods created for supporting human evaluation processes.

2 Hungarian

Hungarian is an agglutinating language, and as such, its morphology is rather complex. Words are often composed of long sequences of morphemes, with agglutination and compounding yielding a huge number of different word forms. For example, while the number of different word tokens in a 20-million-word English corpus is generally below 100,000, the number is above 800,000 in the case of Hungarian. However, the 1:8 ratio does not correspond to the ratio of the number of possible word forms between the two languages: while there are about at most 4–5 possible different inflected forms for an English word, there are about a 1000 for Hungarian, which indicates that a corpus of the same size is much less representative for Hungarian than it is for English [14]. These characteristics often make the direct adaptation of NLP methods developed for English unfeasible. The best performing methods for English often perform significantly worse for Hungarian.

For morphologically rich languages, morphological analysis plays a crucial role in most natural language processing tasks, and the quality of the morphological analyzer used is of great importance. In Hungarian, the morphological behavior of words is also affected by certain semantic features. Proper characterization of semantically restricted morphological constructions is only possible if these features are explicitly listed in the stem database of the analyzer.

The aim of this paper is thus twofold. First, to investigate the performance of neural embedding models applied to a morphologically rich language. Second, to provide a methodology for the automatic derivation of semantic categories relevant from the aspect of morphology.

3 Related Work

The main point of distributional semantics is that the meaning of words is closely related to their use in certain contexts [3]. Traditional models of distributional semantics build word representations by counting words occurring in a fixed-size context of the target word [2].

In contrast, a more recent method for building distributional representations of words is using *word embedding models* the most influential implementation of which is presented in Mikolov et al. [8, 7]. Different implementations of this technique all build continuous vector representations of word meanings from raw corpora. These vectors point to certain locations in the semantic space consistently so that semantically and/or syntactically related words are close to each other, while unrelated ones are more distant. Moreover, it has been shown that vector operations can also be applied to these representations, thus the semantic relatedness of two words can be quantified as the algebraic difference of the two vectors representing these words. Similarly, the meaning of the composition of two words is generally represented well by the sum of two corresponding meaning vectors [9]. One of the main drawback of this method, however, is that by itself it is not able to handle polysemy and homonymy, since one representational vector is built for one lexical element regardless of the number of its different meanings. There are some studies addressing this issue as well by extending the original implementation of word embedding methods [1, 4, 18].

When training embedding models, a fixed-size context of the target word is used, similarly to traditional, discrete distributional models. However, this context representation is used as the input of a neural network. This network is used to predict the target word from the context by using back-propagation and adjusting the weights assigned to the connection between the input neurons (each corresponding to an item in the whole vocabulary) and the projection layer of the network. This weight vector can finally be extracted and used as the embedding vector of the target word. Since similar words are used in similar contexts, these vectors optimized for the context will also be similar for such words. There are two types of neural networks used for this task. One of them is the so called CBOW (continuous bag-of-words) model in which the network is used to predict the target word from the context, while the other model, called skip-gram, is used to predict the context from the target word. For both models, the embedding vectors can be extracted from the middle layer of the network and can be used alike in both cases.

4 Experiments

We built two types of models using the `word2vec`[1] tool, a widely-used framework for creating word embedding representations. This tool implements both models that can be used for building the embedding vectors, however, as the CBOW model has proved to be more efficient for large training corpora, we used this

[1] https://code.google.com/p/word2vec/

model. As a training corpus we used a 3-billion-word raw web-crawled corpus of Hungarian (applying boilerplate removal). In each experiment, the radius of the context window was set to 5 and the number of dimensions to 300.

Then we applied different types of preprocessing to the corpus in order to adapt the method to the agglutinating behavior of Hungarian (or to any other morphologically rich language having a morphological analyzer/tagger at hand).

4.1 The model trained on raw text

First, we built a model from the tokenized but otherwise raw corpus (`SURF`). This model derived different vectors for the different surface forms of the same word. Thus, the various suffixed forms of the same lemma were placed at different locations in the semantic space. As a consequence, this model was able to represent morphological analogies. For example 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'. Table 1 shows some more examples for the list of the most similar forms retrieved for some surface word forms. As it can be seen from the examples, the model represents both semantic and morphosyntactic similarities. For example the top-n list (containing the n most similar words for the target word) for the wordform `kenyerek` 'bread.plur' has similar pastries listed in their plural form (the *-k* ending of all of these words is due to the plural suffix *-k*). The numbers in the lists next to each word are their corpus frequencies.

Even though this model is able to reflect semantic relations to some extent besides morphosyntactic groupings of words, the different surface forms of the same lemma make the model less robust, since the contexts a word is used in are divided between the different surface forms of the same lemma. For example, there are 197 different inflected forms for the lemma `kenyér` 'bread' in the corpus.

4.2 A model built from annotated texts

In the other experiment, we used a morphologically annotated version of the corpus. This was done using the PurePos part-of-speech tagger [15] which also performs lemmatization using morphological analyses generated by the Hungarian Humor morphological analyzer [11, 17, 10]. Each word form in the corpus was represented by two tokens: a lemma token followed by a morphosyntactic tag token `ANA`. Table 2 shows a sentence preprocessed this way.

Since the tags were kept in the actual context of the word they belonged to, the morphosyntactic information carried by the inflections still had a role in determining the embedding vectors. On the other hand, data sparseness was reduced, because the various inflected forms were represented by a single lemma. Table 3 shows some examples of top-n lists generated by this model. While the `SURF` model is often not capable to capture the semantics of rare word forms reliably (e.g. the most similar entries for the word form *Vakkalit* 'Vakkali.Acc'

kenyerek(2270) 'breads'	pirosas(1729) 'reddish'	egerekkel(634) 'with mice'
kiflik(349) 'bagels'	lilás(2476) 'purplish'	patkányokkal(524) 'with rats'
zsemlék(283) 'buns'	rózsaszínes(1638) 'pinkish'	férgekkel(513) 'with worms'
lepények(202) 'pies'	barnás(6463) 'brownish'	majmokkal(606) 'with monkeys'
pogácsák(539) 'scones'	sárgás(7365) 'yellowish'	hangyákkal(343) 'with ants'
pékáruk(771) 'bakery products'	zöldes(5215) 'greenish'	nyulakkal(366) 'with rabbits'
péksütemények(997) 'pastry.pl'	fehéres(2517) 'whitish'	legyekkel(252) 'with flies'
sonkák(613) 'hams'	vöröses(5496) 'reddish'	rágcsálókkal(259) 'with rodents'
tészták(2466) 'pasta.pl'	feketés(1157) 'blackish'	hüllőkkel(241) 'with reptiles'
kalácsok(277) 'cakes'	narancssárgás(429) 'orangish'	pókokkal(436) 'with spiders'
kekszek(1046) 'biscuits'	sárgászöld(723) 'yellowish green'	bogarakkal(425) 'with bugs'

fiaik(1230) 'their sons'	megeszi(7647) 'he eats it'	Vakkalit(5) 'Vakkali.Acc'
lányaik(593) 'their daughters'	eszi(12615) 'he is eating it'	tevedesnek(5) 'as a mıstake'
leányaik(251) 'their daughters'	megenné(563) 'he would eat it'	áfa-jának(7) 'of its VAT'
férjeik(759) 'their husbands'	lenyeli(1862) 'he swallows it'	mot-nak(5) 'mot.Dat'
gyermekeik(12028) 'their children'	megeszik(6433) 'they eat it'	Villanysze(5) 'Electrici(an)'
feleségeik(638) 'their wives'	Megeszi(189) 'He eats it'	oktávtól(5) 'from octave'
gyerekeik(5806) 'their children'	megette(7868) 'he ate it'	Isten-imádat(5) 'worship of God'
asszonyaik(458) 'their wives'	megrágja(477) 'he chews it'	Nagycsajszi(5) 'Big Chick'
gyermekei(31241) 'his children'	megeheti(287) 'he may eat it'	-fontosnak(7) '-as important'
fiak(1523) 'sons'	bekapja(977) 'he swallows it'	tárgykörbôl(5) 'frôm the subject'

Table 1: Similar words in the model created from a raw corpus. Numbers in parentheses show corpus frequency.

are completely unrelated forms in Table 1), the `ANA` model is capable of capturing the semantics of the same lexical items because lemmatization alleviates data sparseness problems and morphosyntactic annotation provides additional grammatical information. The most similar entries of *Vakkali* 'Vakkali' in the `ANA` model *(Ánanda, Avalokitésvara, Dordzse, Babaji, Bodhidharma, Gautama, Mahakásjapa, Maitreya, Bódhidharma)* clearly indicate that this is the name of a Buddhist personality.

4.3 Spelling errors and non-standard word forms

Investigating the models also revealed that among the groups of semantically related words, orthographic variations and misspelled forms of these words also appear. When initiating the retrieval of top-n lists with such non-standard forms, the resulting lists contained words with the same type of errors as the seed word, but semantic similarity was also represented in the ranking of these words. From the preprocessed model, typical error types of the lemmatizer could also be collected. Misspellings and lexical gaps in the morphological analyzer may lead to cases where the guesser in the tagger erroneously tags and lemmatizes words. Lemmas resulting from similar errors are grouped together by the model. E.g. the 'lemmas' *pufidzsek(i)* 'puffy jacket', *rövidnac(i)* 'shorts', *napszemcs(i)* 'sunglasses', *szemcs(i)* 'glasses', *szmöty(i)* 'gunk' etc. all lack the ending *-i*. They

a [Det]	török [Adj]	megszállás [N]	nem [Neg]	feltétlenül [Adv]	jelent [V.Past.3Sg.Def]
the	*Turkish*	*occupation*	*not*	*necessarily*	*mean*
a [Det]	népesség [N]	pusztulás [N.Poss3Sg.Acc]	. [.]		
the	*population*	*destruction*	.		

Table 2: Analyzed version of the Hungarian sentence *A török megszállás nem feltétlenül jelentette a népesség pusztulását.* 'Turkish occupation did not necessarily lead to the destruction of the population.'

kenyér	'bread'	eszik	'eat'	csavargó	'vagabond'
hús(136814)	'meat'	iszik(244247)	'drink'	koldus(15793)	'beggar'
kalács(10658)	'milk loaf'	főz(120634)	'cook'	zsivány(3497)	'rogue'
rizs(31678)	'rice'	csinál(1194585)	'make'	haramia(2024)	'ruffian'
zsemle(6690)	'roll'	megeszik(68347)	'eat'	vadember(2497)	'savage'
pogácsa(11066)	'bisquit'	fogyaszt(160724)	'consume'	csirkefogó(2019)	'scoundrel'
sajt(46660)	'cheese'	etet(43539)	'feed'	szatír(1649)	'satyr'
kifli(9715)	'croissant'	zabál(13699)	'gobble'	útonálló(1942)	'highwayman'
krumpli(37271)	'potato '	megiszik(31002)	'drink'	bandita(6334)	'bandit'
búzakenyér(306)	'wheat bread'	eszeget(3928)	'nibble'	suhanc(4144)	'stripling'
tej(113911)	'milk'	alszik(359268)	'sleep'	vándor(14070)	'wanderer'

Table 3: Similar words in the model created from a annotated corpus. Numbers in parentheses show corpus lemma frequency.

result from the guesser erroneously cutting the ending *-it* from the accusative form of these words. The whole class can be corrected by the same operation, or, as a more permanent solution, all members of the class can easily be added to the lexicon of the morphological analyzer. Similar results can be used to improve the quality of the corpus by correcting these errors in the texts themselves, but also for pinpointing errors in the components of the annotation tool chain (the tokenizer, the lemmatizer or the morphological analyzer) [12]. Another perspective of utilizing this feature of these models is making NLP tools handle OOV items in a more fault tolerant manner by having them annotate unknown words by assigning the annotation of known words that are similar according to the model.

Since the corpus we used was a web-crawled corpus, it also contained a lot of slang and non-standard words coming from user-generated and social media sources. The model works well for these types of texts as well, collecting non-standard words with similar meanings in the top-n lists of such terms. Slang variants *mittomén/mittudomén/mittoménmi/mittudoménmi/nemtommi* 'idunnowhat' are grouped together by the model similarly to representations of laughter *hehehe/hihihi/hahaha/höhö/muhaha/heh/Muhaha/muhahaha/höhöhö.*

4.4 Extracting semantic groups

We used the two models to extract coherent semantic groups from the corpus, which could then be used to enrich the lexical categorization system of a morphological analyzer. Since our goal in this task was to organize words along their semantic similarity, rather than their syntactic behavior, we used the `ANA` model only, i.e. the one trained on the analyzed texts. We created a web application to aid the exploration and visualization of the models and the retrieval of semantically restricted vocabulary. For each category an initial word was selected and the top 200 most similar words were retrieved from the model. Then, the top 200 most similar words were retrieved for items selected by a simple mouse click (taken from the bottom of the previous list). This step was repeated about 10 times. Repeated occurrences were filtered out when retrieving the subsequent lists. The result lists were then merged. Moreover, it was also checked by quick inspection whether the lists did in fact contain mostly relevant items. Those that did not, were deleted by a single click. Throwing these words away, the algorithm was applied again resulting in purer lists. Thus, starting from one word for each category, hundreds or thousands of related words could be retrieved semi-automatically with minimal human interaction that could hardly have been done manually. It was also found that for narrower categories, such as 'materials of clothes', retrieving the top 200 words in each iteration resulted in too much noise, thus in these cases we decreased the size of the top-n list in each iteration to 50.

5 Results

We evaluated the task of semantic categorization by manually counting the number of correct and incorrect words in the given category. However, in order to be able to perform this validation efficiently, the result lists were clustered automatically so that these groups could be reviewed at once. Moreover, the words together with their cluster affiliation were displayed in a two-dimensional plot, providing more visual aid to the human evaluator. Clustering and 2D semantic map visualization was integrated into the web application.

5.1 Clustering

To cluster the lexical elements retrieved from the embedding model, we applied hierarchical clustering. The reason for choosing this type of clustering was based on the argument of [16]. The variety and sophistication of written texts makes the prediction of the number of resulting clusters impossible. However, in a hierarchical clustering, the separation of compact clusters can be performed with regard to the organization of similarities of concept vectors. The input of the clustering algorithm was the set of embedding vectors of candidate words retrieved in the previous step. A complete binary tree was constructed applying Ward's minimum variance method [19] as the clustering criterion, in order to get

small, dense subtrees at the bottom of the hierarchy. However, we did not need the whole hierarchy, but separate, compact groups of terms, i.e. well-separated subtrees of the dendrogram. The most intuitive way of defining the cutting points of the tree is to find large jumps in the clustering levels. To put it more formally, the height of each link in the cluster tree is to be compared with the heights of neighboring links below it up to a certain depth. If this difference is larger than a predefined threshold value, then the link is considered inconsistent, and the tree is cut at that point. Cutting the tree at such points resulted in a list of flat clusters containing more closely related words. The density of these clusters can be set by changing the inconsistency value at which point the subtrees are cut dynamically. This clustering of automatically generated word lists effectively grouped items that did not fit the intended semantic category. As a result, instead of checking hundreds of words individually, only the few clusters had to be signed as correct or incorrect, and this judgment could be applied to all words in the cluster. Only in very few cases did we need to break clusters containing both true and false positives. This method decreased the time needed for manual evaluation drastically.

Table 4 shows some examples of the resulting clusters within each category. The closer relations within a cluster can easily be recognized. For example, in the category of occupations, the abbreviated forms of military ranks formed a separate group, or in the case of languages, different dialects of Hungarian were also collected in a single cluster group, and the other clusters are also of similarly good quality if the words really belonged to the target semantic category. Another type of clusters were those which contained words that were semantically relevant from the aspect of the given task, but were not direct members of the category. For example in the case of languages, geographical names which are modifiers of a language or dialect name but are not language names by themselves, (i.e. non-final elements of multiword language names) were grouped together. The third type of clusters were those that contained words definitely not belonging to the given category. These could then easily be identified and removed manually.

5.2 Visualization

Since the embedding vectors place the lexical elements into a semantic space, it is a common practice to visualize this organization. This is done by transforming the high-dimensional vectorspace to two dimensions by applying the t-sne algorithm [6]. The main point of this method is that it places the words in the two-dimensional space so that the distribution of the pairwise distances of elements is preserved. Thus, the organization of the words can easily be reviewed and outstanding groups can easily be recognized.

When applying this visualization to each semantic category, clustering is also represented in the plot by assigning different colors to different clusters. Thus, not only the distance between individual words, but also the distance between clusters can easily be seen in the resulting figure. Figure 1 shows an example of this visualization.

Occupations
költő író drámaszerző prózaíró novellista színműíró regényíró drámaíró
‘poet’ ‘writer’ ‘drama author’ ‘prosaist’ ‘novelist’ ‘playwright’ ‘novelist’ ‘dramatist’
ökológus entomológus zoológus biológus evolúcióbiológus etológus
‘ecologist’ ‘entomologist’ ‘zoologist’ ‘biologist’ ‘evolutionary biologist’ ‘ethologist’
hidegburkoló tapétázó mázoló szobafestő festő-mázoló szobafestő-mázoló bútorasztalos
‘tiler’ ‘paper hanger’ ‘painter’ ‘housepainter’ ‘painter’ ‘housepainter’ ‘cabinetmaker’
tehénpásztor kecskepásztor birkapásztor fejőnő marhahajcsár tehenész marhapásztor
‘cowherd’ ‘goatherd’ ‘shepherd’ ‘milkmaid’ ‘cattleman’ ‘cowman’ ‘herdsman’
őrm ftörm zls alezr vőrgy szkv ezds hdgy őrgy szds fhdgy
‘Sgt.’ ‘Sgt. Maj.’ ‘WO1’ ‘Lt. Col.’ ‘Maj. Gen.’ ‘Corp.’ ‘Col.’ ‘Lt.’ ‘Maj.’ ‘Capt.’ ‘1Lt.’
Languages
szaúdi kuvaiti szaúd-arábiai jordániai egyiptomi (arab)
‘Saudi’ ‘Kuwaiti’ ‘Saudi Arabian’ ‘Jordanian’ ‘Egyptian (Arabic)’
lengyel cseh bolgár litván román szlovák szlovén horvát
‘Polish’ ‘Czech’ ‘Bulgarian’ ‘Lithuanian’ ‘Romanian’ ‘Slovak’ ‘Slovenian’ ‘Croatian’
osztrák-német német-osztrák elzászi dél-tiroli flamand
‘Austrian-German’ ‘German-Austrian’ ‘Alsatian’ ‘South Tyrolean’ ‘Flemish’
bánsági háromszéki gömöri széki gyimesi felföldi sárközi
Hungarian dialects
Mass nouns
feketeszén kőszén barnaszén lignit feketekőszén barnakőszén
‘black coal’ ‘hard coal’ ‘brown coal’ ‘lignite’ ‘hard coal’ ‘brown coal’
fluorit rutil apatit aragonit kvarc kalcit földpát magnetit limonit
‘fluorite’ ‘rutile’ ‘apatite’ ‘aragonite’ ‘quartz’ ‘calcite’ ‘feldspar’ ‘magnetite’ ‘limonite’
konyhasó kálium-klorid nátriumklorid nátrium-klorid
‘table salt’ ‘potassium chloride’ ‘sodium chloride’ ‘sodium chloride’
Textiles
selyemszatén bélésselyem düsesz shantung
‘silk satin’ ‘silk lining’ ‘duchesse’ ‘shantung’
csipke bársony selyem kelme brokát selyemszövet tafota damaszt batiszt
‘lace’ ‘velvet’ ‘silk’ ‘cloth’ ‘brocade’ ‘serge’ ‘taffeta’ ‘damask’ ‘batiste’

Table 4: Words organized into clusters for four investigated semantic groups

5.3 Quantitative evaluation

Due to the clustering and the visualization applied to the sets of words, the validation of the results became very efficient and easy. This was also due to the parameter settings of the clustering, which resulted in smaller but coherent, rather than larger but mixed clusters. The results of the manual evaluation is shown in Table 5.

We evaluated the categorization method for the following semantic categories: languages, occupations, materials and within that textiles, colors, vehicles, greetings and interjections, and units of measure. We categorized the words (or clusters, if they were homogenous) as correct, erroneous or related. The latter category contained words which did not perfectly fit the original category

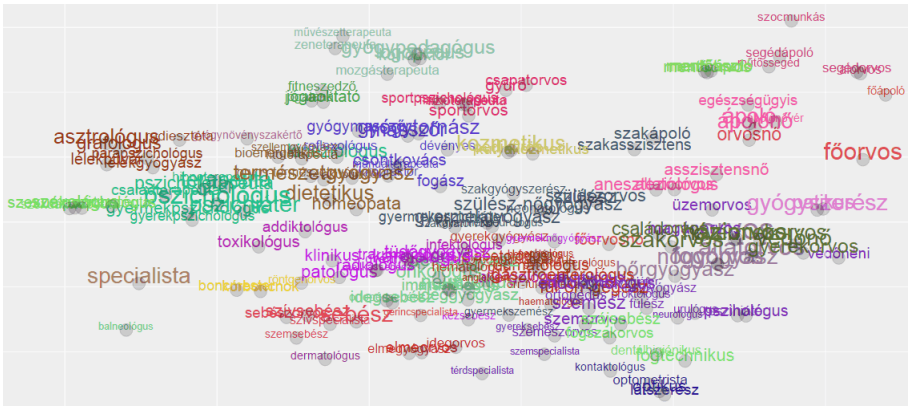

Fig. 1: A fragment of the t-sne visualization of the semantic space of occupations. Distance between words in the diagram is proportional with the semantic similarity of the terms.

	Correct		Erroneous		Related		All
Languages	755	60.69%	98	7.88%	391	31.43%	1244
Occupations	2387	93.32%	134	5.24%	37	1.45%	2558
Materials	1139	84.06%	162	11.96%	54	3.99%	1355
Textiles	120	51.28%	114	48.72%	0	0.00%	234
Colors	870	63.18%	392	28.47%	115	8.35%	1377
Vehicles	1141	72.26%	239	15.14%	199	12.60%	1579
Greetings and interjections	334	24.85%	261	19.42%	749	55.73%	1344
Units of measure	1457	60.08%	909	37.48%	59	2.43%	2425

Table 5: The results of semantic categorization

but were semantically related. In the case of occupations, this category contained words that denote human or humanoid creatures or human roles other than occupation e.g. *srác* 'lad', *öregasszony* 'old woman', *hölgy* 'lady', *albérlő* 'lodger', *élettárs* 'partner', *kobold* 'goblin'. In the case of colors, specific patterns or properties or effects of colors were assigned this category (e.g. *cirmos* 'tabby', *multicolor*, *színehagyott* 'faded'). For vehicles and languages, it contains adjectives which are part of multiword language or vehicle names. The query for greetings resulted in a high number of interjections in addition to greetings, both groups of words exhibiting extreme orthographic variation. In that case, we categorized greetings as correct and interjections as related. The query for units of measures returned also suffixed forms used as heads of adjective phrases, e.g. *5 kilogrammos dinnye* 'a 5 kg melon'. Forms like this are spelled as one word if neither the numeral nor the unit of measure are written using digits or are abbreviated, e.g. *kétórás látogatás* 'a two-hour visit'. Here we categorized forms like this as related. It is evident that, with the exception of textiles, the ratio of correctly categorized items was relatively high. For textiles, on the other hand,

it was extremely easy to check the results because all the incorrect words formed a single distinct cluster that contained only articles made of textiles, clothes, foot gear, home textiles.

6 Conclusion

In this paper, it has been shown that the popular method of neural-network-based embedding models can also be applied to morphologically rich languages like Hungarian, especially if the models are generated from an annotated and lemmatized corpus of a reasonable size. In addition to investigating some of the tasks word embedding models are in general applied to, we demonstrated the applicability of the models for a specific application: expanding the lexicon of a morphological analyzer and extending it with semantic features. We have shown that by applying a semi-automatic method for retrieving words of a certain category and providing further aids for the manual evaluation of these, it has become possible and efficient to assign semantic category labels for words that could not have been done manually. The model has been shown to be capable of pinpointing and categorizing corpus annotation errors as well.

Acknowledgment

We thank Márton Bartók for his help in evaluating automatic semantic categorization results.

References

1. Banea, C., Chen, D., Mihalcea, R., Cardie, C., Wiebe, J.: SimCompass: Using deep learning word embeddings to assess cross-level similarity. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 560–565. Association for Computational Linguistics and Dublin City University, Dublin, Ireland (August 2014)
2. 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). pp. 238–247. Association for Computational Linguistics, Baltimore, Maryland (June 2014)
3. Firth, J.R.: A Synopsis of Linguistic Theory, 1930-1955. Studies in Linguistic Analysis pp. 1–32 (1957)
4. Iacobacci, I., Pilehvar, M.T., Navigli, R.: SensEmbed: Learning sense embeddings for word and relational similarity. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 95–105. Association for Computational Linguistics, Beijing, China (July 2015)
5. Indig, B., Mihăltz, M., Simonyi, A.: Exploiting linked linguistic resources for semantic role labeling. In: 7th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. pp. 140–144. Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu (2015)

6. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-SNE (2008)
7. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
8. 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. pp. 3111–3119 (2013)
9. 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. pp. 746–751 (2013)
10. 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). pp. 1068–1073. European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014), aCL Anthology Identifier: L14-1207
11. Novák, A.: Milyen a jó Humor? [What is good Humor like?]. In: I. Magyar Számítógépes Nyelvészeti Konferencia [First Hungarian conference on computational linguistics]. pp. 138–144. SZTE, Szeged (2003)
12. Novák, A.: Improving corpus annotation quality using word embedding models. Polibits (2016), accepted for publication
13. Novák, A., Siklósi, B., Oravecz, C.: A new integrated open-source morphological analyzer for Hungarian. In: Calzolari, N., Choukri, K., Declerck, T., Grobelnik, M., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA), Portorož, Slovenia (May 2016)
14. Oravecz, C., Dienes, P.: Efficient stochastic part-of-speech tagging for Hungarian. In: LREC. European Language Resources Association (2002)
15. 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). pp. 539–545. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria (2013)
16. Pereira, F., Tishby, N., Lee, L.: Distributional clustering of English words. In: Proceedings of the 31st Annual Meeting on Association for Computational Linguistics. pp. 183–190. ACL '93, Association for Computational Linguistics, Stroudsburg, PA, USA (1993)
17. 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. pp. 261–268. ACL '99, Association for Computational Linguistics, Stroudsburg, PA, USA (1999)
18. Trask, A., Michalak, P., Liu, J.: sense2vec - A fast and accurate method for word sense disambiguation in neural word embeddings. CoRR abs/1511.06388 (2015)
19. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963)