Cross-Lingual Generation and Evaluation of a Wide-Coverage Lexical Semantic Resource

Attila Novák, Borbála Novák
Pázmány Péter Catholic University Faculty of Information Technology and Bionics
MTA-PPKE Hungarian Language Technology Research Group
1083 Budapest, Práter u. 50/a, Hungary
{novak.attila, novak.borbala}@itk.ppke.hu

Abstract

Neural word embedding models trained on sizable corpora have proved to be a very efficient means of representing meaning. However, the abstract vectors representing words and phrases in these models are not interpretable for humans by themselves. In this paper we present the Thing Recognizer, a method that assigns explicit symbolic semantic features from a finite list of terms to words present in an embedding model, making the model interpretable for humans and covering the semantic space by a controlled vocabulary of semantic features. We do this in a cross-lingual manner, applying semantic tags taken form lexical resources in one language (English) to the embedding space of another (Hungarian).

Keywords: semantic lexicon induction, word embedding models, cross-lingual resource generation

1. Introduction

A recently very popular and efficient method for the distributional representation of words is using word embedding (WE) models (Mikolov et al., 2013c). In this paper we present a method that creates the WE model of a large text corpus and inserts the corresponding embedding vectors of a limited set of abstract semantic features into the same space. The embedding vectors for semantic features are built from automatically reorganized lexical resources (that may be in a language different from our target language) and are transformed to the target WE space. Then, a nearest neighbor approach is applied to find the most relevant features for a query word. The assigned features can also be used as a searchable semantic annotation of the original corpus the WE model was created from, because our model assigns semantic features to any (even non-standard/slang or misspelled) word in a text in a language-independent manner, regardless of whether these are present in a lexical resource or not, and whether any such resource is available for the target language. The organization of categories and the way they are actually assigned to words by the algorithm is in accordance with the actual usage of these words as manifested by their distribution in a large corpus.
The method is demonstrated for English and Hungarian, but it can easily be applied to other languages as well.

2. Related Work

WE models have frequently been used to represent word meaning efficiently (Mikolov et al., 2013b; Pennington et al., 2014). There are also approaches that replace WE with sense embedding(Bordes et al., 2012; Neelakantan et al., 2014; Tian et al., 2014; Li and Jurafsky, 2015; Bartunov et al., 2015). Huang et al. (2012) applied clustering algorithms to create single prototype embedding.
Some have tried to match WE's to entities in existing lexical resources, for example to BabelNet entries (Panchenko, 2016) or WordNet synsets (Chen et al., 2014; Agirre et al., 2006). Rothe and Schütze (2015) combines WE vectors to obtain Wordnet synset representations in the original WE space. Labutov and Lipson (2013) also try to take existing WE's and use labeled data to produce WE's in the same space in order to tune or adapt the original representation. Other approaches try to exploit knowledge bases to improve WE's. Yu and Dredze (2014) aim at predicting related words in a knowledge base to WE's. Others compute vector representations of word senses directly from knowledge bases (Bordes et al., 2011; Camacho-Collados et al., 2015).

3. Word Embedding Models for Morphologically Rich Languages

We built WE models for Hungarian, an agglutinative language with complex morphology. In order to incorporate the information encoded in the morphological structure of word forms, full morphological disambiguation was applied to the input words, and the tag sequence following the main PoS tag of each word was detached and included as a separate token following the token consisting of the lemma and the PoS tag in the text. The following example shows the representation of the sentence *Szeretlek, kedvesem.* 'I love you, dear.':

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

Thus, while no information was lost, we managed to improve the quality of the WE model compared to that created from surface word forms in two ways: by assigning a separate representation to lexical items of different part of speech; and by effectively reducing data sparseness problems following from the great variety of rare inflected word forms (Siklósi, 2016).
Although morphological annotation has a less pronounced impact on the quality of the model in English (the language of the lexical resources we used to extract semantic features – see Section 4.), we applied the same method to the English text as well to make the two models compatible

by introducing PoS-based sense distinctions and thus improving the quality of mapping between the models (see Section 5.2.).

For building the WE models, we used the word2vec[1] tool. The Hungarian model was trained on a web-crawled corpus of 3.18 billion tokens (27.49 M token types) that was annotated using the PurePos (Orosz and Novák, 2013) tagger, augmented with the Humor Hungarian morphological analyzer (Novák, 2014; Novák et al., 2016).[2] We trained the English WE model on the English Wikipedia dump of 2.25 billion tokens (8.24 M token types) that was analyzed using Stanford tagger (Toutanova et al., 2003). We created a CBOW model for both languages with the radius of the context window set to 5 and the number of dimensions to 300 and using a token frequency limit of 5. The vocabulary size of the English model was 2,057,592 items and that of the Hungarian one was 2,266,389 items. While only 6 items in the English vocabulary are detached inflectional tags (like [PL] for plural), the Hungarian model contains 2340 such items. These abstract entries representing grammatical morpheme combinations play an important role as context while building the models. The rest are lemmas annotated by their corresponding PoS tag.

4. Lexical Resources

In order to assign semantic labels to the words in the embedding models, we needed some lexical resource to induce the tags from.

A widely used, although quite dated, system of concepts is **Roget's Thesaurus** (Chapman, 1977). Its digitally available third edition contains 990 semantic categories, each further partitioned along five parts-of-speech (noun, verb, adjective, adverb, phrase/interjection). The thesaurus lists a set of related words for each applicable part-of-speech within each category. The original thesaurus includes 91,608 words and multiword expressions. After intersecting it with the vocabulary of the English WE model built from Wikipedia , 51,108 words remained – we lost all MWE's, dated words, and ones with incorrectly marked PoS (see Section 5.1.).

The online version of the digital edition of **Longman Dictionary of Contemporary English** (LDOCE) (Summers, 2005) includes a resource similar to Roget's Thesaurus. However, it contains a much more recent vocabulary and a modern categorization of words. In the online version, words are associated with 213 semantic categories. Part-of-speech is also indicated for each headword. Thus, it could easily be converted to the same format as Roget's Thesaurus, i.e. headwords listed for each part-of-speech in each category. The size of this resource before intersecting with the English WE model was 213 categories and 28,257 example words and multiword entries, which was reduced to 21,546 words after the intersection with the English Wikipedia vocabulary.

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

[2] The annotation generated by this combination of tools contains inflectional features and participles only. The internal structure of compounds and derived words is not explicit in the annotation.

The third resource we used, **4lang**, is also based on LDOCE. The definitions of LDOCE's defining vocabulary were transformed into a formal description (Kornai et al., 2015) illustrated by the following examples:

```
bread: food, FROM/2742 flour, bake MAKE
show: =AGT CAUSE[=DAT LOOK =PAT], communicate
```

We further transformed this format so that we have a similar one to the previous ones. This was achieved by segmenting the formal descriptions into single tokens (by splitting at spaces and brackets) and treating each token as a category label. Then, all words that had the particular token in their definition were listed for that label. This resulted in 1489 category labels and 12,507 words listed for them. 4lang includes some affixes and inflected forms, which are not present in the Wikipedia model, so the intersection resulted in 11,039 words.

We also created another model from 4lang, in which we did not segment predicates with more than one argument into further parts, so e.g. HAS[four.(legs)] remained an atomic feature. Further processing of this model, to which we refer as **4lang2** in the paper, was identical to that of the 4lang model. The first four columns of Table 1 summarize the main characteristics of the resources, while Table 2 shows some examples from each resource.

Resource	Original			After ∩ & clustering		
	cats	words	w/c	cats	words	w/c
LDOCE	213	28257	132.66	3069	21546	7.02
Roget's	3077	91608	29.77	7066	51108	7.23
4lang	1489	12507	8.39	2249	11039	4.91
4lang2	4172	12507	2.99	4256	11039	2.59

Table 1: Characteristics of the three lexical resources (number of different category labels, number of words and the average number of words per category; before and after intersection with the English embedding model and clustering).

One of the most popular semantic resources for English is **WordNet** (Fellbaum, 1998; Miller, 1995). However, WordNet has been criticized for its too high granularity at the bottom level and its generality at the top level (Brown, 2008). Selecting an appropriate set of concepts from WordNet that could be used as semantic features is far from trivial. There is a high level categorization into which WordNet synsets are organized ("supersenses"), and these could be used as features similarly to the ones derived from the resources mentioned before. However, there are only 45 supersenses, which seems to be an extremely low-grained categorization to be useful for practical purposes. Due to these problems, although we consider using WordNet in the future both as a resource and as a possible benchmark, we did not use it in the experiments presented in this paper.

5. Method

The goal of this research was to create a tool that is able to assign semantic features to words, even if the target word is not included in any semantic lexicon or if such a lexicon does not even exist in the given language. Thus, two problems had to be handled: assigning features and, if needed, bridging the language gap.

Resource	Category	Example words in the original resource
ROGET	Mean_N	medium#NN generality#NN neutrality#NN middle_state#NN median#NN golden_mean#NN middle#NN etc.
ROGET	Rotundity_ADJ	spherical#JJ cylindric#JJ round_as_an_apple#JJ bell_shaped#JJ spheroidal#JJ conical#JJ globated#JJ etc.
LDOCE	Cooking	allspice#NN bake#VB barbecue#VB baste#VB blanch#VB boil#VB bottle#VB bouillon_cube#NN etc.
LDOCE	Mythology	centaur#NN chimera#NN Cyclops#NN deity#NN demigod#NN faun#NN god#NN griffin#NN gryphon#NN etc.
4LANG	food	sandwich#NN, fat#NN, bread#NN, pepper#NN, meal#NN, fork#NN, egg#NN, bowl#NN, salt#NN etc.
4LANG	=DAT	say#VB, show#VB, allow#VB, swear#VB, grateful#ADV, let#VB, teach#VB, give#VB, help#VB etc.
4LANG2	PART_OF.body	body#NN, tongue#NN, back#NN, neck#NN, shoulder#NN, bone#NN, skin#NN, wrist#NN, buttock#NN etc.
4LANG2	=AGT.HAS.mouth	swallow#VB, suck#VB, eat#VB, drink#VB

Table 2: Examples from each resource after transforming them to the same format

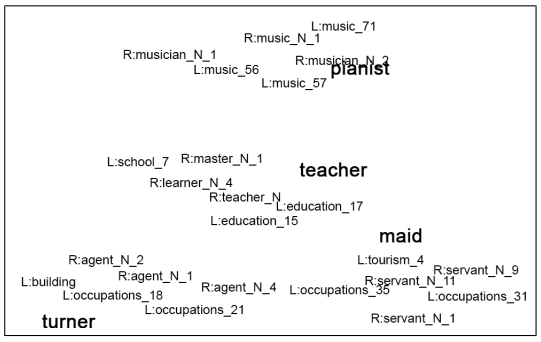

Figure 1: The 3 nearest features assigned to the words *pianist, teacher, turner, maid* from the LDOCE and Roget's models arranged in semantic space

5.1. Semantic Feature Space

As described in Section 4., we used three lexical resources in this experiment using the category labels in these lexicons as semantic features. However, some categories were too broad and the set of words listed for them was too heterogeneous. To handle this problem, a hierarchical agglomerative clustering algorithm was applied to the set of words in those categories that contained at least five words (for details of the clustering algorithm, see (Siklósi, 2016)). Each cluster was then labeled with the original category label and a numeric index. Since the clustering algorithm used the distance between the embedding vectors of words trained from the English Wikipedia corpus, only words present in the Wikipedia model could be used from the original resources. How this intersection and the clustering of words affected the representations in each lexical resource is shown in Table 1.

We used a simple but effective method for representing each semantic feature in the same semantic space as that of the English PoS-tagged WE model: we assigned the average of the embedding vectors of clustered example words to each indexed semantic label. To find the relevant features for a query word tagged with its appropriate part-of-speech, its representational vector is retrieved from the WE model and its nearest neighbors are taken from each feature model. Figure 1 shows how four words (*pianist, teacher, turner, maid*) and the 3 nearest features assigned to them from the LDOCE and Roget's models are organized in semantic space.

5.2. Cross-Lingual Mapping of the Models

It has been shown that WE spaces can effectively be mapped across languages. One mapping method is to use a word-aligned bilingual parallel corpus to build an embedding model that contains vector representations of words in both languages (Luong et al., 2015). We applied another approach instead, where the projection is achieved by learning a piecewise linear transformation based on a seed dictionary, through which a monolingual WE space can be mapped to another monolingual space (Mikolov et al., 2013a). The transformation maps each word vector in the source language space to a point in the vicinity of the vector of its translation in the target language space.

We used a subset of the 4lang dictionary (built from the defining vocabulary of LDOCE) containing 3477 English-Hungarian word pairs as the seed dictionary to calculate the transformation matrix. We used pairs where both the English and the Hungarian word had a frequency over 10000 in the two corpora. Manual evaluation of the transformation on an additional 100 words resulted in 0.38 precision for the first-ranked translation and precision=0.69/0.81 for the first 5/10 top-ranked translations (indicating whether a correct translation of the target word was found in the set of the first five/ten most similar words in the transformed space). We used this transformation matrix to map the English semantic label vectors to the Hungarian WE space. Then, the same nearest neighbor algorithm could be applied to the query word as in the case of searching the English semantic space. This made it possible to input a Hungarian word as a query to our system and receive semantic features based on originally English resources without the expensive and labor-intensive task of translating them. Moreover, since instead of exact matching, nearest neighbors are searched for, out-of-vocabulary words (with respect to the original lexical resources) can also be assigned semantic labels.

6. Experiments and Results

When looking at the output of the models, we found that even though the LDOCE features seemed to be the most meaningful, the Roget's, 4lang and 4lang2 models also turned out to be useful. E.g. adjectives have a much richer categorization in Roget's than what we obtain from the LDOCE model. Since LDOCE and Roget's seemed to perform well in complementary regions, we decided to unify these two models (ROLD).

We carried out two kinds of quantitative analysis of the performance of our model. First, we checked the robustness of the model by performing a sanity check on the original English resources. In the other scenario, we selected 280 words randomly from a predefined list of Hungarian words in which each word was assigned to one of 28 semantic domains (e.g. food, vehicles, locations, occupations, etc.) and manually checked the accuracy of the semantic features assigned to these words by each model.

6.1. Sanity Check

For each word present in the original 4lang dictionary, we calculated how many of the semantic features present in the original definition were retrieved among the top N features returned by the model (feature recall, R_f) and the percentage of words for which all features were retrieved (word recall, R_w). The results are shown in Table 3 as a function of N. Recall was also calculated ignoring words having more than N features ($R_w(poss)$) and features over the N limit ($R_f(poss)$). As no definition contained more than 10 terms, $R_w(poss)$ is identical to R_w and $R_f(poss)$ is identical to R_f for $N \geq 10$. The last column of the table shows the mean reciprocal rank of features (terms) present in the original definitions. Reciprocal rank is calculated as $i/Rank$ for the i^{th} feature returned by the model that is also present in the original definition, it is zero if no valid feature was retrieved. MRR is calculated as the average of the reciprocal rank of all expected features retrieved for all words.

	N	R_w	$R_w(poss)$	R_f	$R_f(poss)$	**MRR**
4lang	1	0.1508	0.8504	0.2694	0.9455	0.9455
	5	0.5472	0.6574	0.7614	0.8445	0.9586
	10	0.7049	0.7080	0.8756	0.8785	0.9237
	20	0.8187	0.8187	0.9316	0.9316	0.8922
4lang2	1	0.4411	0.8818	0.5079	0.9266	0.9266
	5	0.8688	0.8775	0.9138	0.9226	0.9456
	10	0.9339	0.9339	0.9597	0.9597	0.9276
	20	0.9648	0.9648	0.9793	0.9793	0.9163
ROLD	1	0.3354	0.3590	0.7421	0.8426	0.8426
	5	0.6557	0.7482	0.7017	0.8079	0.9080
	10	0.7433	0.8349	0.7481	0.8419	0.8877
	20	0.8117	0.8896	0.8118	0.8897	0.8645

Table 3: Performance (recall) of the three models for English tested on the original resources.

6.2. Standard Language Use

After the sanity check, we tested our system on standard Hungarian. In order to do this, we collected groups of words belonging to different semantic categories. These categories were defined manually and the test words were collected by a semi-automatic algorithm as described in (Siklósi, 2016). Finally, each group was manually checked resulting in 28 groups containing 39,050 words altogether. We randomly selected 10 words from each group, and the top 10 semantic features were generated using the models 4lang, 4lang2 and ROLD. The list of randomly selected words also included misspelled and very rare words. Features that were partitioned and indexed when building the models (see Section 5.1.) were joined after lookup. Two annotators checked the generated semantic feature sets, and marked each feature that was inappropriate for the given lexical item (e.g. `HAS.horn` for *vízimadár* 'water fowl'). Cases when the given lexical group is in the domain of the given feature (e.g. the domain of `HAS.horn` is animals) and completely inappropriate features (e.g feature `dig` for *csűr* 'barn' in the buildings group) were not differentiated: they were all simply marked wrong. Inter-annotator agreement was found to be substantial (Cohen's kappa=0.734). Results of the evaluation are shown in Table 4[3]. The table shows semantic feature accuracy (acc: the ratio of correctly assigned features) in each category for each model. We also automatically computed feature "domain accuracy" (d-acc): here we ignored feature assignment errors where the same feature was marked adequate for another test word in the same domain. The table also shows the number of different features (#F) each model assigned to the test words in each domain, and the number of features that were marked wrong for *any* of the test words in the given domain (#B). The overall feature accuracy of the 4lang-derived models was nearly 75%, while the combined ROLD model achieved over 80% feature accuracy. The feature space of the ROLD model is less fine-grained in some domains (e.g.food or clothing) than that derived from 4lang definitions (this is indicated by the lower number of different features assigned by the ROLD model) and this results in higher accuracy. Note that the domain accuracy of 4lang features is much higher than feature accuracy, it is about 90%. The worst average accuracy was obtained on colors: lists of things having specific colors or patterns and the high number of color terms themselves generated too much noise.

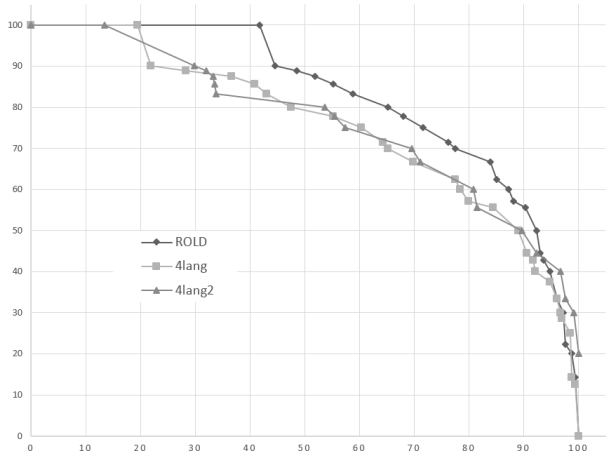

Figure 2: The distribution of feature precision for the three models ROLD, 4lang and 4lang2.

Figure 2 shows the distribution of the precision of features per word. The ROLD model assigned only appropriate features to 42% of the 290 test words, and precision was over 70% for over 75% of the words. 4lang and 4lang2 had 100% precision for 20% and 13.4% of the test words, respectively. All models had over 50% precision for about 90% of the words. The precision of 4lang2 was over 20% for all test words.

6.3. Proper Names and Non-standard Language

The WE models our method is based on also reflect world knowledge as represented in the corpus from which they are generated from. This enables our model to assign features to proper names of various types, such as names of people, institutions, fictional creatures, or even abbreviations as shown in Table 5. In the names section of the table, some famous people are shown, one of them is Hungarian (*Béla Bartók*, a Hungarian composer). It can be seen

[3]Due to length limits, we included only selected categories in the extended abstract.

	4lang					4lang2					ROLD				
Group name	acc	d-acc	#F	#B	most frq. features	acc	d-acc	#F	#B	most frq. features	acc	d-acc	#F	#B	most frq. features
Units of measure	62.22	74.44	55	26	unit measure	63.64	74.75	59	27	unit measure	70.00	81.67	34	13	Measurement Computers
Electronics	86.36	90.91	38	7	machine device	82.00	88.00	43	13	equipment machine	78.26	85.51	35	13	Technology Recording
Diseases	88.71	94.62	37	13	bad body	87.88	95.45	46	10	bad bad(situation)	94.23	98.08	24	2	Illness Disease_N
Animals	69.23	94.87	35	20	wild animal	67.19	94.79	35	18	animal HAS.wings	84.21	93.86	24	9	Animal_N Animals
Kitchen utensils	79.49	91.03	31	14	instrument contain	76.00	89.00	39	18	food.IN metal	92.31	92.31	24	3	Receptacle_N Daily_life
Food	57.14	95.24	26	14	food COOK	64.52	96.77	23	10	food material	97.44	97.44	7	1	Food Food_N
Vehicles	84.71	95.29	34	10	vehicle engine	74.55	92.73	45	15	vehicle ‹HAS.engine›	75.00	88.16	27	7	Journey_N Vehicle_N
Clothes	68.35	96.20	16	6	WEAR garment	71.00	87.00	34	14	garment cloth	100.00	100.00	12	0	Clothing_N Clothes
Disciplines	77.91	86.05	51	14	science educate	76.00	87.00	52	15	science knowledge	88.24	92.94	48	10	Education Knowledge_N
Water	87.64	98.88	24	7	water valley	84.00	99.00	29	8	land ON.earth	98.25	100.00	18	1	Geography Geology
Geographic areas	83.91	96.55	31	8	land valley	76.53	96.94	29	10	land natural	87.14	94.29	27	7	Geography Geology
Natural events	85.37	91.46	50	11	wind atmosphere	85.86	91.92	57	13	weather water	72.29	80.72	45	19	Meteorology Nature
Mountains, hills	79.89	89.66	32	11	hill land	80.30	92.42	27	13	hill mountain	93.86	98.25	17	4	Geography Nature
Cities	85.29	94.12	27	6	city place	82.61	92.75	25	8	city place	87.84	90.54	40	8	Abode_N Geography
Locations	85.94	93.75	31	7	country land	87.93	93.10	22	5	country land	81.54	86.15	41	10	Geography Government_N
Buildings	76.47	85.29	30	12	building place	86.00	92.00	39	11	place building	87.50	94.44	30	9	Abode_N Buildings
Groups of humans	90.00	96.47	55	9	group purpose	85.57	90.21	52	15	institution structure	85.71	92.86	55	14	Organizations Receptacle_N
Human relationship	74.73	95.60	43	14	=POSS KNOW	78.00	98.00	39	13	HAS.parent companion	82.35	88.24	48	12	Auxiliary_N Friend_N
Athletes	59.09	81.82	34	19	PLAY game	59.00	74.00	44	19	sport person	71.95	84.15	27	14	Strength_N Other_sports
Occupations	71.76	76.47	52	21	profession science	75.00	80.00	51	20	profession person	84.38	90.63	39	10	Occupations Scholar_N
Time	79.41	86.76	42	13	sunset monday	57.50	71.25	50	29	period ON.earth	46.43	69.64	31	21	Chronology Celebration_N
Events	60.53	76.32	39	20	invite holiday	76.53	92.86	35	16	FOR.pleasure period	81.40	94.19	35	14	Amusement_N Leisure
Colors	45.98	89.66	20	11	colour shade	40.86	84.95	18	14	colour light	53.33	68.33	28	16	Colours Color_Adj
Attributes of humans	67.76	95.39	48	18	stupid good	75.26	93.16	60	22	strange heavy	85.47	88.83	100	24	Love_Adj Badness_Adj
Attributes of food	71.43	86.81	53	16	taste COOK	82.65	89.80	46	10	material sweet	100.00	100.00	23	0	Food Food_dish
Verbs of movement	71.25	88.75	24	10	rush long	52.00	61.00	29	19	go rush	46.39	61.86	49	30	Velocity_Vb Journey_Vb
Verbs of free-time	50.59	75.29	29	21	relax <person>	60.00	77.00	46	24	relax lack(work)	59.79	75.26	61	31	Outdoor Endearment_Vb
Verbs of decay	69.88	84.34	43	16	after CAUSE	55.67	74.23	54	26	slip die	68.48	75.00	69	25	Death_Vb Chemistry
All	**74.74**	**90.07**	**564**	**266**		**73.86**	**88.34**	**584**	**295**		**80.36**	**87.93**	**561**	**252**	

Table 4: Performance of the models 4lang, 4lang2 and ROLD on test words from different semantic groups. **acc**: feature accuracy, **d-acc**: domain accuracy of features, **#F**: different features, **#B**: features marked wrong at least once.

Bartók	4L: music art *poem *poet *poetry WRITE sound *text musician 4L2: art *poem *poet music HAS.rhythm entertainment sound sequence *text MAKE.beautiful RL: Music Music_N Performing
Obama	4L: country government politician @United_States state LEAD *place president republic 4L2: country politician @United_States country.HAS place MAKE.law state *@Soviet_Union politics RL: Officials Government_N Government Politics_N Authority_N Director_N Council_N
MTA	4L: institution group society *president *republic educate science purpose *person people 4L2: institution society group educate science HAS.purpose study structure people RL: Occupations Education *Receptacle_N College *Geology Skill_N Organizations
ELTE	4L: educate institution study student degree science numbers atom *GIVE 4L2: educate institution study science *name *part knowledge public *system RL: College Education Knowledge_N School_Adj Language_N
PPKE	4L: educate institution science group study student degree society *sleeve @Catholic_Church 4L2: educate study institution science knowledge group religion *system job HAS.purpose RL: College School_Adj Education Occupations School

Table 5: Examples of features returned for proper names and abbreviations of names of institutions from the models

that each person is assigned features that provide information about them. Thus, the model can be queried even for names one is not familiar with, and relevant features will be provided. This also holds for names with lower frequency in the corpus, as long as the name itself is unique.

Table 5 also contains the abbreviated name of some organizations. *ELTE* is for Eötvös Loránd University, while *PPKE* is for Pázmány Péter Catholic University. While both of them are educational institutions, *ELTE* is a state university, but *PPKE* is catholic, and this difference is reflected by the set of features assigned to them in addition to their relation to science and education.

The same applies to slang terms, including many short diminutive forms. These are abundant in the web-crawled corpus, mainly coming from often heated discussions in user comments and fora, and many of them have strong emotional connotations. These are neatly reflected by the semantic tags assigned to them in addition to the ones reflecting the basic meaning of the term, e.g. 'Deceiver', 'Obstinacy', 'Ignorance', 'Thief', 'Crime', 'Politics' 'Race relations' 'Psychology, Psychiatry', 'stupid', 'criminal' in addition to 'person' for derogative terms like *nyugger* 'pensioner', *proli* 'proletarian', *bolsi* 'bolshevik' or *cigó* 'Gypsy'.

7. Conclusions

We have shown that the meaning implicitly represented in word embedding models can be transformed into a set of symbolic features that can be used as semantic annotation. This can also be done across languages, thus relevant semantic tags can be assigned to words in a language that lacks appropriate semantic resources. Despite its simplicity, our system, the Thing Recognizer, performs this surprisingly efficiently also for names and words that cannot be expected to be included in manually created lexical semantic resources.

Acknowledgments

This research has been implemented with support provided by grants FK125217 and PD125216 of the National Research, Development and Innovation Office of Hungary financed under the FK17 and PD17 funding schemes.

8. Bibliographical References

Agirre, E., Martínez, D., de Lacalle, O. L., and Soroa, A. (2006). Evaluating and optimizing the parameters of an unsupervised graph-based wsd algorithm. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs-1, pages 89–96, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bartunov, S., Kondrashkin, D., Osokin, A., and Vetrov, D. (2015). Breaking sticks and ambiguities with adaptive skip-gram. *arXiv preprint arXiv:1502.07257*.

Bordes, A., Weston, J., Collobert, R., and Bengio, Y. (2011). Learning structured embeddings of knowledge bases. In *AAAI*.

Bordes, A., Glorot, X., Weston, J., and Bengio, Y. (2012). Joint learning of words and meaning representations for open-text semantic parsing. In *In Proceedings of 15th International Conference on Artificial Intelligence and Statistics*.

Brown, S. W. (2008). Choosing sense distinctions for wsd: Psycholinguistic evidence. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 249–252, Stroudsburg, PA, USA. Association for Computational Linguistics.

Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2015). A unified multilingual semantic representation of concepts. 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)*, pages 741–751, Beijing, China, July. Association for Computational Linguistics.

Chapman, R. (1977). *Roget's International Thesaurus*. Harper Colophon Books. Crowell.

Chen, X., Liu, Z., and Sun, M. (2014). A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035, Doha, Qatar, October. Association for Computational Linguistics.

Christiane Fellbaum, editor. (1998). *WordNet: an electronic lexical database*. MIT Press.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kornai, A., Ács, J., Makrai, M., Nemeskey, D. M., Pajkossy, K., and Recski, G. (2015). Competence in lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 165–175, Denver, Colorado, June. Association for Computational Linguistics.

Labutov, I. and Lipson, H. (2013). Re-embedding words. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 489–493, Sofia, Bulgaria, August. Association for Computational Linguistics.

Li, J. and Jurafsky, D. (2015). Do multi-sense embeddings improve natural language understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732, Lisbon, Portugal, September. Association for Computational Linguistics.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Bilingual word representations with monolingual quality in mind. In *NAACL Workshop on Vector Space Modeling for NLP*, Denver, United States.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013a). Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). 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.*, pages 3111–3119.

Mikolov, T., Yih, W., and Zweig, G. (2013c). 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*, pages 746–751.

Miller, G. A. (1995). Wordnet: A lexical database for English. *COMMUNICATIONS OF THE ACM*, 38:39–41.

Neelakantan, A., Shankar, J., Passos, A., and McCallum, A. (2014). Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar, October. Association for Computational Linguistics.

Novák, A., Siklósi, B., and Oravecz, C. (2016). A new integrated open-source morphological analyzer for Hungarian. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Novák, A. (2014). A new form of Humor – mapping constraint-based computational morphologies to a finite-state representation. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

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

Panchenko, A. (2016). Best of both worlds: Making word sense embeddings interpretable. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Rothe, S. and Schütze, H. (2015). Autoextend: Extending word embeddings to embeddings for synsets and lexemes. 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)*, pages 1793–1803, Beijing, China, July. Association for Computational Linguistics.

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

Summers, D. (2005). *Longman Dictionary of Contemporary English*. Longman Dictionary of Contemporary English Series. Longman.

Tian, F., Dai, H., Bian, J., Gao, B., Zhang, R., Chen, E., and Liu, T.-Y. (2014). A probabilistic model for learning multi-prototype word embeddings. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 151–160, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yu, M. and Dredze, M. (2014). Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 545–550, Baltimore, Maryland, June. Association for Computational Linguistics.