Semantic Regularity of Derivational Relations

Pavel Šmerk

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
smerk@fi.muni.cz

Abstract. The paper presents a very preliminary attempts to employ the *word2vec* tool to evaluate semantic regularity within particular derivational relations in the data of the *Derivancze* derivational analyzer of Czech.

Keywords: derivational morphology, semantics of the derivational relations, word2vec

1 Introduction

In [1] we (together with Karel Pala) presented *Derivancze*, derivational analyzer for Czech. We gave reasons, why we prefer not to include all derivational relations, but to limit our data only to those relations which are semantically transparent and regular. But we had no objective measure which could assure us that our relations actually are semantically regular.

In this paper we present an attempt to employ *word2vec* [2,3,4] tool to evaluate the semantic regularity of derivational relations in our data. In the following section we shortly describe derivational relations of *Derivancze* and the data of the analyzer. In the next section we present our experiments and results, namely with some "averaged concept", which should represent the semantics of a given derivational relation. Finally we discuss some of the possible future improvements.

2 Derivational Relations and Data of *Derivancze*

The following description of the derivational relations is taken from our previous paper [1] (refer to it for more information):

- `k1verb`, `k2pas`, `k2proc`, `k2rakt`, `k2rpas`, and `k2ucel` from verbs, where:
  - `k1verb` derives nouns describing process, action or state denoted by the verb (*kropit–kropení*; *sprinkle–sprinkling*),
  - `k2pas` and `k2rpas` are passive participle and past passive adjectival participle, i. e. two forms of adjectives which describe the patient or object of the action (*kropit–kropen / kropený*; *sprinkle–sprinkled*),

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2015, pp. 87–92, 2015.

  - `k2proc` derives present active adjectival participles, i. e. adjectives describing a subject doing the action (*kropit–kropící*; *sprinkle–sprinkling* (man)),
  - `k2rakt` are past active adjectival participles, i. e. adjectives describing subjects which have completed the action (*pokropit–pokropivší*; *sprinkle–who has springled st.*), and
  - `k2ucel` derives adjectives which describe an object used for the action (*kropit–kropicí*; *sprinkle–sprinkling* (machine)),
- verb → agent noun relation `k1ag` (*bádat–badatel*; *research–researcher*),
- adjective → name of the property relation `k1prop` (*rychlý–rychlost*; *fast–speed*),
- adjective → adverb relation `k6a` (*dobrý–dobře*; *good–well*),
- noun → possessive adjective relation `k2pos` (*otec–otcův*; *father–father's*),
- noun → relational adjective relation `k2rel` (*virus–virový*; *virus–viral/virus*), semantically perhaps the most heterogenous relation among the Derivancze relations,
- relations `k1f`, `k1jmf`, and `k1jmr` express changes in gramatical gender:
  - `k1f` derives feminines from general masculines (*doktor–doktorka*; *doctor*$_{\text{MASC}}$ *–doctor*$_{\text{FEM}}$),
  - `k1jmf` derives feminine forms of surnames (*Novák–Nováková*), and
  - `k1jmr` derives family forms of surnames (*Novák–Novákovi*) — it should be noted that `k1f` and `k1jmf` cannot be joined because of names of nationalities, which can also act as surnames, but the derived forms differ (*Rus–Ruska* X *Rusová*, i. e. *Russian*$_{\text{FEM}}$ X *Mrs. Rus*),
- area or city → inhabitant name relation `k1obyv` (*Kanada–Kanaďan*; *Canada–Canadian*), formally the most heterogenous relation in Derivancze,
- noun → deminutive relation `k1dem` (*dům–domek*; *house–little house*).

The Table 1 shows a distribution of the derivational pairs in *Derivancze* data according to the derivational relation. The row `var` is the semantic equivalence (see [1] for details). The numbers are without the pairs in which both members are the same[1]. The column **total** is the total number of pairs in our data. The columns **N**+ are numbers of pairs whose both members occur in the corpus CzTenTen [5] (ca 5.3 billion tokens) with a frequency at least N.

3 Experiments

One of the possible benchmarks for evaluating the word vectors produced by the *word2vec* tool is complementing triplets of form for instance

Greece Athens Norway ?
Kazakhstan Astana Zimbabwe ?
...

[1] For instance, the passive participle (`k2pas`) of the verb *přejet* is *přejet*.

Table 1. Distribution of derivational pairs according to relation.

Relation	total	1+	10+	50+	1000+
k1ag	703	400	400	248	89
k1dem	6334	4122	4122	2983	1218
k1f	3196	1800	1800	1253	413
k1jmf	2211	1893	1893	1735	442
k1jmr	2207	1112	1112	427	31
k1obyv	261	215	215	168	91
k1prop	9902	5708	5708	3775	1110
k1verb	35723	15848	15848	11186	3906
k2pas	34779	7334	7334	4567	1249
k2pos	30936	7110	7110	3853	539
k2proc	15765	5017	5017	3462	1054
k2rakt	18106	369	369	101	9
k2rel	23518	15343	15343	10972	4045
k2rpas	35015	12377	12377	8177	2772
k2ucel	1672	1402	1402	1081	370
k6a	45108	11399	11399	6615	2108
var	1136	592	592	390	112
total	266572	92041	92041	60993	19558

(according to [2]). These triplets represent questions like "what is the word that is similar to Norway in the same sense as Athens is similar to Greece?" and the model should predict *Oslo* and *Harare*, respectively. The method, evaluation data and results are thoroughly described in many publications, e. g. [2,3,4].

In fact, we are not interested in some abstract similarity of *Greece* and *Athens* but these words are only particular representants of some common question "what is the capital city of ...". Interestingly, the results are much better if instead of a particular representant we use an average of more such examples.

3.1 "Averaged Concept"

As an illustration we present results of a very trivial experiment. We took the first 50 most frequent pair from `k1f` relation except the pair *muž/žena* (*man/woman*) and for these pairs and the word *muž* we evaluated questions in the aforementioned sense, so the answer should always be *žena*. We do the tests on models computed from 120 million, 1 billion, and 5.3 billion tokens from the corpus CzTenTen (the last is the whole corpus). Then we computed the average of these first 50 pairs and evaluated the question with these averages. The results are in the Table 2.

The numbers are always the cosine distance and can be interpreted as a measure of similarity (for further details again refer to [2,3,4]). For the first

Table 2. Experiment with an "averaged concept" representing `k1f`.

	120M		1G		5.3G	
avg	0.617013		0.744693		0.775647	
min	0.434191	občan/občanka	0.509271	občan/občanka	0.541018	občan/občanka
max	0.729597	otec/matka	0.857678	táta/máma	0.889231	táta/máma
1	**0.799218**	žena	**0.877726**	žena	**0.890839**	žena
2	0.648836	dívka	0.754475	dívka	0.762988	dívka
3	0.503029	ženský	0.610383	ženský	0.623256	chlapec
4	0.471445	otrokyně	0.595448	chlapec	0.621017	ženský
5	0.467925	matka	0.573623	matka	0.600329	mužův
6	0.467344	chlapec	0.564603	děvče	0.595389	dívenka
7	0.467089	milenec	0.555519	mladík	0.591686	děvče
8	0.466913	mladík	0.550952	mužův	0.582920	matka
9	0.462479	tmavovlasý	0.546331	partnerka	0.575861	partnerka
10	0.461867	chlap	0.541432	dívenka	0.572923	družka

50 pairs we present an average, minimum[2] and maximum[3] distance. The following ten rows are the top ten best answers for the "averaged" question. It can be easily seen that the pattern is the same for all three sizes of training data: the "averaged" right answer is always better than the result of any single pair, always much better than the average of the single pairs and always much better than any other possible answer — it should be added that *dívka* (as well as *dívenka* and *děvče*) is *young woman*, which means that even the second best answer is rather acceptable. The interesting is that all these effects are stronger for smaller data.

3.2 Evaluation of Derivational Relations

It suggests that for evaluation of semantic regularity of derivational relations we should try to find the concepts represented by the particular relations and check all pairs against that concept. As a very preliminary experiment we took the first 50 most frequent pairs of each derivational relation[4], compute the vector of this averaged concept and compare it with the first 150 most frequent pairs. It was evaluated on the whole corpus CzTenTen. The results are in the left half of the Table 3: the TOP10 column is the number of pairs where the correct answer was among the first 10 words according to the model (the whole vocabulary of the model has almost 1.5 million entries), the TOP1 column is the number of pairs where the correct answer was the first, the rank column is the average rank of the correct answer, and the distance is the average distance.

[2] In Czech *občanka* is not only a feminine form of *citizen* but also *identity card*.

[3] *otec* is *father*, *matka* is *mother*, *táta* is *daddy*, and *máma* is *mummy*.

[4] We do not explore `k1jmf` and `k1jmr` because they represent relations between proper names (surnames) and the automatic lemmatisation of the corpus is unfortunately of a very low quality for these words.

Table 3. Towards the concept of particular derivational relations.

	TOP10	TOP1	rank	distance	TOP10	TOP1	rank	distance
k1ag	53	6	4.42	0.595907	47	3	3.81	0.613398
k1dem	88	39	1.57	0.709737	85	40	1.59	0.718124
k1f	123	95	0.59	0.783001	123	98	0.56	0.786888
k1obyv	133	85	1.06	0.699675	133	86	1.00	0.703276
k1prop	88	30	1.69	0.654496	89	35	1.74	0.659931
k1verb	130	59	1.63	0.708456	127	56	1.83	0.712785
k2pas	134	99	0.59	0.828821	131	78	0.81	0.792059
k2pos	102	61	1.10	0.744492	95	51	1.36	0.740843
k2proc	119	79	1.14	0.738209	118	76	1.14	0.738124
k2rakt	44	3	3.75	0.563547	47	3	3.83	0.566550
k2rel	122	62	1.12	0.717147	123	56	1.57	0.713637
k2rpas	127	38	1.98	0.719895	124	40	2.06	0.719619
k2ucel	50	1	4.78	0.632999	52	1	4.71	0.636890
k6a	118	53	1.59	0.627307	116	65	1.31	0.635010

The second step was an attempt to refine the concept by selecting the best 50 pairs according to the initial concept. The results are in the right half of the Table 3. As there were many oddities in the data we calculated the average ranks and distances (in both halves of the Table 3!) only from the "TOP10" pairs, because it is OK for a wrong pair to be less similar to the refined concept, but the limit 10 is entirely arbitrary for the present.

4 Conclusion and Future Work

The results in the previous section show that word vectors computed by *word2vec* are useful for checking the semantic consistency of the derivational relations, at least to some extent. The number of pairs in TOP10 or even TOP1 is consistently rather high and many of the pairs which were not "successful" were either clearly wrong, e. g. *plat* (*salary*), *plátek* (*slice*) for `k1dem`, or with some very irregular semantics, e.g. *věřit* (*believe*), *věřitel* (*creditor*) for `k1ag` where the second is derived from the first, but it is very uncommon to say that *věřitel věří* (*creditor believes*).

In the future we plan to improve lemmatisation of the corpus, as the experiments revealed many systematic errors which obviously degrade the results. Both *word2vec* and the setup of the experiments have many parameters which were set almost arbitrarily and we will try to fine tune them. The refinement of the average concept was successful for some of the relations, but not for all, and although it was successful in the average, the improvement was much smaller than expected, thus we believe there is also a space for a further improvement. Then, of course, the refinement iteration should continue up to some fix point. The hugest amount of manual work will be demanded by removal of wrong pairs from the data. After all, we will be able to offer not only

the derivational pairs alone, but also some additional information how close is the semantics of the particular pair to the average semantics of the respective derivational relation.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

References

1. Pala, K., Šmerk, P.: Derivancze — Derivational Analyzer of Czech. In Král, P., Matoušek, V., eds.: Text, Speech, and Dialogue. Volume 9302 of Lecture Notes in Computer Science., Springer (2015) 515–523
2. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
3. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q., eds.: 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
4. Mikolov, T., Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In Vanderwende, L., III, H.D., Kirchhoff, K., eds.: 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, The Association for Computational Linguistics (2013) 746–751
5. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family. International Conference on Corpus Linguistics, Lancaster (2013)