Typology of Adjectives Benchmark for Compositional Distributional Models

Daria Ryzhova[1], **Maria Kyuseva**[1], **Denis Paperno**[2]
[1]NRU Higher School of Economics, [2]University of Trento
[1]School of Linguistics – Staraya Basmannaya Ulitsa 21/4, 105066 Moscow, Russia;
[2]CIMeC – Center for Mind/Brain Sciences Palazzo Fedrigotti – corso Bettini 31, 38068 Rovereto (TN), Italy
daria.ryzhova@mail.ru, mkyuseva@gmail.com, denis.paperno@unitn.it

Abstract

In this paper we present a novel application of compositional distributional semantic models (CDSMs): prediction of lexical typology. The paper introduces the notion of typological closeness, which is a novel rigorous formalization of semantic similarity based on comparison of multilingual data. Starting from the Moscow Database of Qualitative Features for adjective typology, we create four datasets of typological closeness, on which we test a range of distributional semantic models. We show that, on the one hand, vector representations of phrases based on data from one language can be used to predict how words within the phrase translate into different languages, and, on the other hand, that typological data can serve as a semantic benchmark for distributional models. We find that compositional distributional models, especially parametric ones, perform way above non-compositional alternatives on the task.
Keywords: compositional distributional models, adjectives, lexical typology

1. Introduction

It is well-known that meanings of words can vary from language to language. For example, there is no Russian adjective that could be used to translate the English adjective *sharp* in all contexts: *sharp edge*, *sharp needle*, *sharp turn*, *sharp image*, *sharp mind*, etc., and similarly no English adjective translates the Russian *ostryj* 'sharp' in all contexts, which include among others *ostryj rog* 'sharp horn' and *ostryj perec* 'hot pepper'. Can independently motivated semantic models predict such variation of lexical semantics across languages? For example, do existing semantic models provide clues suggesting that the adjective *sharp* in many but not all languages expresses the meaning of 'spicy'? This question has no obvious answer: semantic models are usually built for one particular language and as such are not meant to account for linguistic variation. However, it has been argued that patterns like 'sharp'→'spicy' are not arbitrary but are motivated by an inherent semantic basis, which is common for different languages (Rakhilina and Reznikova, 2013; Koptjevskaja-Tamm et al., 2015). If linguists are right about this, an adequate model of natural language semantics should be able to predict such patterns. In search for the answer to our question, we evaluate several (compositional) distributional semantic models against data on lexical semantic typology of adjectives.

1.1. Distributional Semantic Models

Distributional semantic models (DSMs) are able to predict many kinds of phenomena, including word similarity, lexical ambiguity, and semantic priming, or even to pass TOEFL synonymy and analogy tests (Landauer and Dumais, 1997; Griffiths et al., 2007; Turney and Pantel, 2010). Compositional distributional semantic models (CDSMs) attempt to extend vector representations to arbitrary phrases (Mitchell and Lapata, 2008; Baroni and Zamparelli, 2010), up to full sentences (Socher et al., 2011; Socher et al., 2012; Pham et al., 2013; Paperno et al., 2014).

CDSMs have been applied to tasks ranging from paraphrase detection (Socher et al., 2011) to phrase plausibility (Vecchi et al., 2011) to improving syntactic parsing (Lazaridou et al., 2013) to sentiment analysis (Socher et al., 2012). However, one might wonder how much properly semantic content CDSMs actually capture beyond the syntactic and distributional properties of words and phrases. Some of the tasks to which (C)DSMs have been applied are syntactic in nature, and others might not require deep semantic analysis. For example, very shallow cues like lexical overlap and keyword matching are often surprisingly informative for tasks like sentence similarity and sentiment analysis. Predicting semantically motivated trends in lexical typology is a novel, theoretically principled way to assess to what extent CDSMs capture compositional meanings.

1.2. Lexical Typology

Linguistic typology studies the diversity of natural languages on different levels: diversity of syntactic structures, grammatical categories, anaphora patterns, etc. Lexical typology is a subfield of linguistic typology that focuses on how the meanings of open-class lexical items vary from one natural language to another (Berlin and Kay, 1969; Wierzbicka, 1996; Majid et al., 2007). Semantic fields that have been examined most extensively include color terms, kinship terms, spacial relations, and motion verbs.

Typological data used in this paper is taken from the Moscow Database of Qualitative Features, based on the Moscow tradition of lexical typology (Rakhilina and Reznikova, 2013; Kyuseva et al., 2013). The approach taken in creating the Database rests upon the idea of the universal set of minimal lexical meanings, independently supported by recently published work (Youn et al., 2016). For example, there are three minimal meanings in the direct sense zone of the semantic field 'sharp': 'sharp cutting edge', 'sharp piercing point', and 'sharp form'. Languages differ in the way they cluster these three meanings within lexemes. In English, the adjective *sharp* covers them all, in Japanese they are distributed between *surudoi* ('cutting edge' and 'piercing point') and *togatta* ('form'), the contrast in French is between *tranchant* ('cutting edge') and *pointu* ('piercing point' and 'form'), and so on. Traditional

typological research requires manual analysis of a minimum of 10–12 languages to more or less reliably converge on the universal set of minimal meanings for a given semantic field.

The set of minimal meanings for every semantic zone is reflected on the distribution of the lexemes belonging to the field. Indeed, the English *sharp* combines with nouns denoting various kinds of objects: instruments with a cutting edge (*knife* or *sword*) and a piercing point (*needle* or *arrow*), as well as objects with a narrowing end (for example, *nose*), while the Japanese *togatta* occurs in combinations with the nouns from the third group only, and the French *pointu* with nouns from the second and the third groups.

2. Toward a Dataset of Typological Closeness

Up to now, (C)DSMs have been applied to tasks in the languages they were trained on. If, however, (C)DSMs capture meaning as the invariant properties of text preserved in translation to any language, they have to be applied to tasks involving translation equivalents; in particular, any viable model of semantics should account for how the expression of the same meanings *varies crosslinguistically*.

In typology, cross-linguistic variation is usually characterized qualitatively. To use lexical semantic variation for quantitative evaluation of semantic models, we define the notion of *typological closeness*, which can be measured numerically:

> **Typological closeness** of two word usages w, w' is a measure of how likely an arbitrary lexical item of an arbitrary human language that covers one of them is to cover both.

For example, the uses of the adjective in *sharp stick* and *sharp nose* (as opposed to *sharp knife*) are typologically close: translation equivalents for the phrases *sharp stick* and *sharp nose* in other languages will typically use the same adjective (French *bâton pointu*, *nez pointu*), but the most natural translation of *sharp knife* will often use a distinct adjective (*couteau tranchant*).

The notion of **typological closeness** just illustrated is a novel formalization of semantic similarity, to which (C)DSMs are immediately applicable. The fact that two given situations are often described with the same lexeme in many different languages provides an objective clue about their semantic closeness. Without denying the utility of traditional similarity judgments, we propose typology as a novel application for compositional distributional semantic models, and when typological data are available, they do have natural advantages as a semantic benchmark. Unlike human judgments, typological closeness is not affected by factors external to semantics such as word frequency, collocation frequency, phonetic similarity, morphological relatedness, or by mere misunderstanding of the task.

3. Creating the Dataset from Lexical Typology of Adjectives

We build four test sets of typological closeness as defined above, to which we apply a variety of compositional distributional semantic models.

We extracted the typological data to construct the test dataset from the Database containing typological questionnaires for physical qualities ('thick', 'sharp', 'wet', etc.) filled with data from 30 languages that belong to diverse areal groups and language families. Typological questionnaires are tables with adjective translation equivalents as columns and contexts in which the adjectives can occur as rows (as we have shown above, the adjective's context can usually be represented by the noun it modifies). An intersection of a row and a column is filled with '1' if the adjective can occur in the context and with '0' if it cannot (see examples in Table 1).

The list of contexts includes nouns presupposing both direct and figurative usages of the adjective in question (e.g. 'sharp knife' vs. 'sharp contrast'). Rows of the typological questionnaire are vectors **w** whose dimensions are translation equivalents from all the languages studied:

> $\mathbf{w}_i = 1$ iff w is covered by the lexeme l_i of some language in the database.

For example, a typological vector for '... knife', obtained from Table 1, is [1,1,0,0,0].

As a measure of *typological closeness* between corresponding adjective meanings in context, we use the *cosine* of two typological vectors. For example, the cosine between the typological vectors for '[sharp] stick' and '[sharp] nose' equals 0.82, while for '[sharp] stick' and '[sharp] knife' it is 0.56.

With cosine as a numeric measure of typological closeness, we created test benchmarks using data from two unrelated semantic fields ('sharp' and 'smooth'). The Database of Qualitative Features contains 33 lexemes that cover the field 'sharp' in 15 languages and 32 lexemes from the field 'smooth' (9 languages).

For every possible pair of rows of the typological questionnaires ('[sharp] stick' and '[sharp] nose', '[sharp] stick' and '[sharp] knife', '[sharp] knife' and '[sharp] nose', and so on) we computed the value of typological closeness, obtaining four subsets of test data:

- semantic field 'sharp', direct and figurative meanings (9019 pairs of rows)
- semantic field 'sharp', only direct meanings (528 pairs)
- semantic field 'smooth', direct and figurative meanings (1992 pairs)
- semantic field 'smooth', only direct meanings (561 pairs)

For constructing these datasets, we adopted the distinction of direct and figurative adjective uses as annotated in the Database.

4. Evaluation of Distributional Models

4.1. Model Construction

To assess the viability of our approach, we tested several known compositional distributional models on the dataset described above. Evaluation was based on distributional

	English *sharp*	Japanese *surudoi*	Japanese *togatta*	French *tranchant*	French *pointu*
... knife	1	1	0	1	0
... sword	1	1	0	1	0
... needle	1	1	0	0	1
... nose	1	0	1	0	1

Table 1: A fragment of the typological questionnaire for the field 'sharp'. Columns correspond to individual adjectives. For illustrative purposes, we show here only a selection of columns. Some languages mentioned in the table have more adjectives in the semantic field 'sharp'.

vectors of Russian noun phrases matched with the corresponding slots of the questionnaires. For example, the slot '... knife' for the field 'sharp' was matched with the phrase *ostryj nozh* 'sharp knife', substituting the gap with the relevant Russian adjective. For each pair of adjective-noun phrases in the questionnaire, we calculated the cosine of the corresponding CDSM vectors, and compared it to typological closeness of the phrases in question.

The vectors were built from cooccurrence counts of words and phrases with the 10K most frequent nouns, verbs, adjectives and adverbs as contexts, collected in a fixed window of ± 5 content words based on data from three Russian corpora (total size about 1.44 bln tokens). The phrase vectors used to train the CDSMs were derived from the cooccurrence profiles of relevant phrases. Thus, the CDSMs under evaluation were trained on distributional data and tested on typological data, without supervision on the task. The parameters of the distributional models were:

- corpus size and genre
  We used three Russian corpora of different size and genres:
  1. the main subcorpus of the Russian National Corpus (RNC, http://ruscorpora.ru), a highly balanced collection of texts, a total size of about 220 million tokens;
  2. the RNC subcorpus of newspaper and journal texts (about 200 million tokens);
  3. ruWaC (http://corpus.leeds.ac.uk/internet.html), which consists primarily of Internet text (about 1 billion tokens).
- vector processing
  Cooccurrence counts were converted into vectors with the different parameter values available in the DISSECT toolkit (http://clic.cimec.unitn.it/composes/toolkit/, (Dinu et al., 2013): weighting (none, PPMI, PLMI, EPMI and PLOG) and dimensionality reduction (none or SVD to 300 dimensions).
- vector type: observed vs. composed
  We tested both vectors produced from the noun phrase's cooccurrence patterns and compositionally derived vectors, obtained with the help of several known compositional models (also available in the DISSECT toolkit): additive, weighted additive, multiplicative, dilation, lexical function, and practical lexical function (Paperno et al., 2014) models.

For the evaluation metric, we use Pearson's correlation coefficient between (C)DSM cosine and typological closeness. Pearson's r was chosen over Spearman's rank-based ρ as the figure of merit because our data contain a significant number of ties between datapoints. The correlation value can be interpreted as a measure of semantic models' informativeness about the typology of a given semantic field.

For comparison, we also used a non-compositional noun-only model (i.e. using the distributional vector of the noun for the full "adjective + noun" phrase). For example, the vector of the Russian noun *nozh* 'knife' rather than the noun phrase *ostryj nozh* 'sharp knife' represented the row '... knife'. The noun-only baseline was expected to be quite strong, because different senses of an adjective usually apply to different types of objects, and appear with different classes of nouns (see, for example, (Rakhilina, 2008). However, we found the performance of the noun-only baseline to be quite low (0.274 under the best parameter settings).

4.2. Evaluation Results

Table 2 contains the results of our experiments, reporting the scores of the models with the best parameter settings and, for comparison, several lower-performing ones. All models produced statistically significant correlation with typological data ($p < 0.01$). The highest correlation coefficients are given in bold. The main result is that the methodology developed here is viable: we can predict to a large degree the cross-linguistic properties of adjective usage from monolingual compositional vectors.

In both semantic fields typological closeness and DSM similarity are much more correlated if we consider only direct meanings. This supports a widely accepted point of view that figurative word meanings are less structured than direct meanings, and therefore less predictable cross-linguistically.

With the exception of corpus size, all the DSM parameters explored affect model performance.

Corpus size. According to a popular view, increase in corpus size should lead to better vector quality. For example, it was found that word vectors extracted from the web-crawled ukWaC corpus outperform on a variety of semantic tasks similar vectors from the smaller but better balanced BNC (Bullinaria and Levy, 2012). Our experiment involved

	corpus	vectors	composition model	sharp	sharp:direct	smooth	smooth:direct
1	RNC	none	noun only	0.092	0.12	0.121	0.196
2	RNC	ppmi	noun only	0.139	0.167	0.237	0.21
3	RNC	ppmi,SVD	noun only	0.167	0.268	0.274	0.244
4	RNC	ppmi	none	0.097	0.194	0.134	0.154
5	RNC	ppmi	Additive	0.36	0.654	0.589	0.74
6	RNC	ppmi	Multiplicative	0.253	0.421	0.585	0.7
7	RNC	ppmi	Dilation	0.19	0.222	0.379	0.443
8	RNC	ppmi	Dilation w/ Training	0.207	0.35	0.249	0.313
9	RNC	ppmi, SVD	LexFunc	0.112	0.336	0.263	0.349
10	RNC	ppmi, SVD	LexFunc, Ridge	0.116	0.345	0.443	0.703
11	RNC	ppmi, SVD	PrLexFunc	0.389	0.765	0.444	0.931
12	RNC	ppmi, SVD	PrLexFunc, Ridge	0.39	**0.766**	0.449	**0.946**
13	RNC	none	WeightedAdd	0.443	0.754	0.589	0.849
14	RNC	plog	WeightedAdd	0.387	0.76	0.477	0.765
15	RNC	epmi	WeightedAdd	**0.462**	0.763	0.59	0.865
16	RNC	ppmi	WeightedAdd	0.42	0.764	**0.604**	0.905
17	RNC	plmi	WeightedAdd	0.443	0.762	0.603	0.791
18	all	ppmi	WeightedAdd	0.418	0.764	0.564	0.899
19	all	plmi	WeightedAdd	0.438	0.763	0.549	0.712
20	RNC	ppmi, SVD	Additive	0.269	0.443	0.404	0.566
21	RNC	ppmi, SVD	Dilation w/ Training	0.388	0.766	0.448	0.936
22	RNC	ppmi, SVD	WeightedAdd	0.388	0.717	0.421	0.682
23	RNC	ppmi, SVD	Dilation	0.231	0.519	0.374	0.512
24	RNC	ppmi, SVD	Multiplicative	0.062	0.41	0.194	0.228

Table 2: CDSMs evaluation results

the Russian National Corpus which is comparable to BNC and the ruWaC corpus built in the same way as ukWaC. Surprisingly, in our evaluation even a sevenfold corpus size increase didn't improve the results (compare rows 16-19 in the table). A well-balanced corpus of 200 mln tokens sufficed to collect word vectors and phrase data for high quality composition training.

Vector space parameters. In the course of the experiment different weighting schemes as well as vector normalization and dimensionality reduction were applied. EPMI and PPMI weightings proved to be particularly effective (compare rows 13-17). Dimensionality reduction improves the results of certain composition models but not others. Some of the highest correlation coefficients were obtained with the model that applied SVD to the PPMI matrix, in line with the findings by (Baroni et al., 2014).

Composition. Despite the fact that a 200 mln word corpus proved to be large enough to construct high-quality word vectors, even a 1.44 bln token corpus doesn't suffice to collect quality phrase vectors directly, treating the whole phrase like an indivisible unit (see row 4 in the table). We evaluated several composition models, and all of them significantly improve the result above the noncompositional baseline.

5. Discussion

This paper presents a novel application of distributional semantic models: they can predict lexical typology. Our approach might be beneficial for typology and computational semantics alike, and may help develop useful tools for multilingual lexicography even for resource-poor languages.

The notion of typological closeness that we defined allowed us to create a new benchmark of semantic relatedness, targeted at the semantic aspects of phrase similarity relying on translation equivalent statistics in many languages. We tested an array of compositional distributional models on this typological similarity dataset. On the one hand, the evaluation confirmed that CDSMs capture language-independent semantics. Indeed, semantic vectors built for a single language were able to predict in an unsupervised setting typological patterns that hold across languages. On the other hand, the evaluation showed that our test set is a valid semantic benchmark. All vector space models make statistically significant predictions, and the model parameters, which had not been optimized for the task, affect the performance in expected and meaningful ways: weighting, composition and specifically parametric composition models perform better than unweighted, non-compositional or nonparametric alternatives, respectively. In future research, we are going to extend our methodology to include other parts of speech such as nouns, verbs, and adverbs, hence another types of diagnostic contexts.

6. Acknowledgements

We thank Marco Baroni, Germán Kruszewski, Tal Linzen, Marco Marelli, Ekaterina Rakhilina, and members of the CLIC lab for valuable input. This research is supported by the Russian Foundation for Basic Research, grant 14-06-00343 A, and by the ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES).

7. Bibliographical References

Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, pages 1183–1193, Boston, MA.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). 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)*, pages 238–247. Association for Computational Linguistics.

Berlin, B. and Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. University of California Press.

Bullinaria, J. A. and Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, 44:890–907.

Dinu, G., Pham, N. T., and Baroni, M. (2013). Dissect: Distributional semantics composition toolkit. In *Proceedings of ACL (System Demonstrations)*, pages 31–36, Sofia, Bulgaria.

Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, 114(2):211.

Koptjevskaja-Tamm, M., Rakhilina, E., and Vanhove, M., (2015). *The semantics of lexical typology*, pages 434–454. Routledge.

Kyuseva, M., Reznikova, T., and Ryzhova, D. (2013). Tipologicheskaya baza dannyh adjektivnoy leksiki [a typologically oriented database of qualitative features]. In *Kompjuternaya lingvistika i intellectual'nyye tehnologii [Computational Linguistics and Intellectual Technologies]*, volume 1, pages 419–430. RSUH, Moscow.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Lazaridou, A., Vecchi, E., and Baroni, M. (2013). Fish transporters and miracle homes: How compositional distributional semantics can help NP parsing. In *Proceedings of EMNLP*, Seattle, WA. In press.

Majid, A., Bowerman, M., Van Staden, M., and Boster, J. S. (2007). The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cognitive Linguistics*, 18(2):133–152.

Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. In *Proceedings of ACL*, pages 236–244, Columbus, OH.

Paperno, D., Pham, N. T., and Baroni, M. (2014). A practical and linguistically-motivated approach to compositional distributional semantics. In *Proceedings of ACL*, pages 90–99, Baltimore, Maryland.

Pham, N. T., Bernardi, R., Zhang, Y.-Z., and Baroni, M. (2013). Sentence paraphrase detection: When determiners and word order make the difference. In *Proceedings of the Towards a Formal Distributional Semantics Workshop at IWCS 2013*, pages 21–29, Potsdam, Germany.

Rakhilina, E. and Reznikova, T. (2013). Frejmovy podhod k leksicheskoj tipologii [frame approach to lexical typology]. *Voprosy yazykoznania*, 2:3–31.

Rakhilina, E. (2008). *Kognitivny analiz predmetnyh imen: semantika i sochetaemost' [Cognitive Analysis of Physical Names: Semantics and Combinability]*. Russkie slovari, Moscow, 2nd edition.

Socher, R., Huang, E., Pennin, J., Ng, A., and Manning, C. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of NIPS*, pages 801–809, Granada, Spain.

Socher, R., Huval, B., Manning, C., and Ng, A. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP*, pages 1201–1211, Jeju Island, Korea.

Turney, P. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Vecchi, E. M., Baroni, M., and Zamparelli, R. (2011). (Linear) maps of the impossible: Capturing semantic anomalies in distributional space. In *Proceedings of the ACL Workshop on Distributional Semantics and Compositionality*, pages 1–9, Portland, OR.

Wierzbicka, A. (1996). *Semantics: Primes and Universals*. Oxford University Press, Oxford.

Youn, H., Suttond, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., Croft, W., and Bhattacharya, T. (2016). On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences of the United States of America*, 113:1766–1771.