Building and Evaluating a Distributional Memory for Croatian

Jan Šnajder* **Sebastian Padó†** **Željko Agić‡**
*University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
†Heidelberg University, Institut für Computerlinguistik
69120 Heidelberg, Germany
‡University of Zagreb, Faculty of Humanities and Social Sciences
Ivana Lučića 3, 10000 Zagreb, Croatia
jan.snajder@fer.hr pado@cl.uni-heidelberg.de zagic@ffzg.hr

Abstract

We report on the first structured distributional semantic model for Croatian, DM.HR. It is constructed after the model of the English Distributional Memory (Baroni and Lenci, 2010), from a dependency-parsed Croatian web corpus, and covers about 2M lemmas. We give details on the linguistic processing and the design principles. An evaluation shows state-of-the-art performance on a semantic similarity task with particularly good performance on nouns. The resource is freely available.

1 Introduction

Most current work in lexical semantics is based on the *Distributional Hypothesis* (Harris, 1954), which posits a correlation between the degree of words' semantic similarity and the similarity of the contexts in which they occur. Using this hypothesis, word meaning representations can be extracted from large corpora. Words are typically represented as vectors whose dimensions correspond to context features. The vector similarities, which are interpreted as semantic similarities, are used in numerous applications (Turney and Pantel, 2010).

Most vector spaces in current use are either *word-based* (co-occurrence defined by surface window, context words as dimensions) or *syntax-based* (co-occurrence defined syntactically, syntactic objects as dimensions). Syntax-based models have several desirable properties. First, they are model to fine-grained types of semantic similarity such as predicate-argument plausibility (Erk et al., 2010). Second, they are more versatile – Baroni and Lenci (2010) have presented a generic framework, the Distributional Memory (DM), which is applicable to a wide range of tasks beyond word similarity. Third, they avoid the "syntactic assumption" inherent in word-based models, namely that context words are relevant iff they are in an n-word window around the target. This property is particularly relevant for free word order languages with many long distance dependencies and non-projective structure (Kübler et al., 2009). Their obvious problem, of course, is that they require a large parsed corpus.

In this paper, we describe the construction of a Distributional Memory for Croatian (DM.HR), a free word order language. To do so, we parse hrWaC (Ljubešić and Erjavec, 2011), a 1.2B-token Croatian web corpus. We evaluate DM.HR on a synonym choice task, where it outperforms the standard bag-of-word model for nouns and verbs.

2 Related Work

Vector space semantic models have been applied to a number of Slavic languages, including Bulgarian (Nakov, 2001a), Czech (Smrž and Rychlý, 2001), Polish (Piasecki, 2009; Broda et al., 2008; Broda and Piasecki, 2008), and Russian (Nakov, 2001b; Mitrofanova et al., 2007). Previous work on distributional semantic models for Croatian dealt with similarity prediction (Ljubešić et al., 2008; Janković et al., 2011) and synonym detection (Karan et al., 2012), however using only word-based and not syntactic-based models.

So far the only DM for a language other than English is the German DM.DE by Padó and Utt (2012), who describe the process of building DM.DE and the evaluation on a synonym choice task. Our work is similar, though each language has its own challenges. Croatian, like other Slavic languages, has rich inflectional morphology and free word order, which lead to errors in linguistic processing and affect the quality of the DM.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 784–789,
Sofia, Bulgaria, August 4-9 2013.

3 Distributional Memory

DM represents co-occurrence information in a general, non-task-specific manner, as a tensor, i.e., a three-dimensional matrix, of weighted *word-link-word* tuples. Each tuple is mapped onto a number by scoring function $\sigma\colon W \times L \times W \to \mathbb{R}^+$, that reflects the strength of the association. When a particular task is selected, a vector space for this task can be generated from the tensor by matricization. Regarding the examples from Section 1, synonym discovery would use a *word by link-word* space ($W \times LW$), which contains vectors for words w represented by pairs $\langle l, w\rangle$ of a link and a context word. Analogy discovery would use a *word-word by link* space ($WW \times L$), which represents word pairs $\langle w_1, w_2\rangle$ by vectors over links l.

The links can be chosen to model any relation of interest between words. However, as noted by Padó and Utt (2012), dependency relations are the most obvious choice. Baroni and Lenci (2010) introduce three dependency-based DM variants: DepDM, LexDM, and TypeDM. DepDM uses links that correspond to dependency relations, with subcategorization for subject (*subj_tr* and *subj_intr*) and object (*obj* and *iobj*). Furthermore, all prepositions are lexicalized into links (e.g., ⟨*sun*, ***on***, *Sunday*⟩). Finally, the tensor is symmetrized: for each tuple $\langle w_1, l, w_2\rangle$, its inverse $\langle w_2, l^{-1}, w_1\rangle$ is included. The other two variants are more complex: LexDM uses more lexicalized links, encoding, e.g., lexical material between the words, while TypeDM extends LexDM with a scoring function based on lexical variability.

Following the work of Padó and Utt (2012), we build a DepDM variant for DM.HR. Although Baroni and Lenci (2010) show that TypeDM can outperform the other two variants, DepDM often performs at a comparable level, while being much simpler to build and more efficient to compute.

4 Building DM.HR

To build DM.HR, we need to collect co-occurrence counts from a corpus. Since no sufficiently large suitable corpus exists for Croatian, we first explain how we preprocessed, tagged, and parsed the data.

Corpus and preprocessing. We adopted hrWaC, the 1.2B-token Croatian web corpus (Ljubešić and Erjavec, 2011), as starting point. hrWaC was built with the aim of obtaining a cleaner-than-usual web corpus. To this end, a conservative boilerplate removal procedure was used; Ljubešić and Erjavec (2011) report a precision of 97.9% and a recall of 70.7%. Nonetheless, our inspection revealed that, apart from the unavoidable spelling and grammatical errors, hrWaC still contains non-textual content (e.g., code snippets and formatting structure), encoding errors, and foreign-language content. As this severely affects linguistic processing, we additionally filtered the corpus.

First, we removed from hrWaC the content crawled from main discussion forum and blog websites. This content is highly ungrammatical and contains a lot of non-diacriticized text, typical for user-generated content. This step alone removed one third of the data. We processed the remaining content with a tokenizer and a sentence segmenter based on regular expressions, obtaining 66M sentences. Next, we applied a series of heuristic filters at the document- and sentence-level. At the document level, we discard all documents (1) whose length is below a specified threshold, (2) contain no diacritics, (3) contain no words from a list of frequent Croatian words, or (4) contain a single word from lists of distinctive foreign-language words (for Serbian). The last two steps serve to eliminate foreign-language content. In particular, the last step serves to filter out the text in Serbian, which at the sentence-level is difficult to automatically discriminate from Croatian. At the sentence-level, we discard sentences that are (1) shorter than a specified threshold, (2) contain non-standard symbols, (3) contain non-diacriticized Croatian words, or (4) contain too many foreign words from a list of foreign-language words (for English and Slovene). The last step filters out specifically the sentences in English and Slovene, as we found that these often occur mixed with text in Croatian. The final filtered version of hrWaC contains 51M sentences and 1.2B tokens. The corpus is freely available for download, along with a more detailed description of the preprocessing steps.[1]

Tagging, lemmatization, and parsing. For morphosyntactic (MSD) tagging, lemmatization, and dependency parsing of hrWaC, we use freely available tools with models trained on the new SETimes Corpus of Croatian (SETIMES.HR), based on the Croatian part of the SETimes parallel corpus.[2] SETIMES.HR and the derived tools are prototypes

[1] http://takelab.fer.hr/data

[2] http://www.nljubesic.net/resources/corpora/setimes/

	SETIMES.HR	Wikipedia
HunPos (POS only)	97.1	94.1
HunPos (full MSD)	87.7	81.5
CST lemmatizer	97.7	96.5
MSTParser	77.5	68.8

Table 1: Tagging, lemmatization, and parsing accuracy

that are about to be released as parts of another work. Here we give a general description and a re-evaluation that we consider relevant for building DM.HR.

SETIMES.HR consists of 90K tokens and 4K sentences, manually lemmatized and MSD-tagged according to Multext East v4 tagset (Erjavec, 2012), with the help of the Croatian Lemmatization Server (Tadić, 2005). It is used also as a basis for a novel formalism for syntactic annotation and dependency parsing of Croatian (Agić and Merkler, 2013).

On the basis of previous evaluation for Croatian (Agić et al., 2008; Agić et al., 2009; Agić, 2012) and availability and licensing considerations, we chose HunPos tagger (Halácsy et al., 2007), CST lemmatizer (Ingason et al., 2008), and MSTParser (McDonald et al., 2006) to process hrWaC. We evaluated the tools on 100-sentence test sets from SETIMES.HR and Wikipedia; performance on Wikipedia should be indicative of the performance on a cross-domain dataset, such as hrWaC. In Table 1 we show lemmatization and tagging accuracy, as well as dependency parsing accuracy in terms of labeled attachment score (LAS). The results show that lemmatization, tagging and parsing accuracy improves on the state of the art for Croatian. The SETIMES.HR dependency parsing models are publicly available.[3]

Syntactic patterns. We collect the co-occurrence counts of tuples using a set of syntactic patterns. The patterns effectively define the link types, and hence the dimensions of the semantic space. Similar to previous work, we use two sorts of links: unlexicalized and lexicalized.

For unlexicalized links, we use ten syntactic patterns. These correspond to the main dependency relations produced by our parser: *Pred* for predicates, *Atr* for attributes, *Adv* for adverbs, *Atv* for verbal complements, *Obj* for objects, *Prep* for prepositions, and *Pnom* for nominal predicates. We subcategorized the subject relation into *Sub_tr* (subjects of transitive verbs) and *Sub_intr* (subject of intransitive verbs). The motivation for this is better modeling of verb semantics by capturing diathesis alternations. In particular, for many Croatian verbs reflexivization introduces a meaning shift, e.g., *predati* (*to hand in/out*) vs. *predati se* (*to surrender*). With subject subcategorization, reflexive and irreflexive readings will have different tensor representations; e.g., ⟨*student*, *Subj_tr*, *zadaća*⟩ (⟨*student*, *Subj_tr*, *homework*⟩) vs. ⟨*trupe*, *Subj_intr*, *napadač*⟩ (⟨*troops*, *Subj_intr*, *invadors*⟩). Finally, similar to Padó and Utt (2012), we use *Verb* as an underspecified link between subjects and objects linked by non-auxiliary verbs.

[3] http://zeljko.agic.me/resources/

Link	P (%)	R (%)	F_1 (%)
Unlexicalized			
Adv	57.3	52.7	54.9
Atr	85.0	89.3	87.1
Atv	75.3	70.9	73.1
Obj	71.4	71.7	71.5
Pnom	55.7	50.8	53.1
Pred	81.8	70.6	75.8
Prep	50.0	28.6	36.4
Sb_tr	67.8	73.8	70.7
Sb_intr	64.5	64.8	64.7
Verb	61.6	73.6	67.1
Lexicalized			
Prepositions	67.2	67.9	67.5
Verbs	61.6	73.6	67.1
All links	73.7	75.5	74.6

Table 2: Tuple extraction performance on SETIMES.HR

For lexicalized links, we use two more extraction patterns for prepositions and verbs. Prepositions are directly lexicalized as links; e.g., ⟨*mjesto*, ***na***, *sunce*⟩ (⟨*place*, ***on***, *sun*⟩). The same holds for non-auxiliary verbs linking subjects to objects; e.g., ⟨*država*, ***kupiti***, *količina*⟩ (⟨*state*, ***buy***, *amount*⟩).

Tuple extraction and scoring. The overall quality of the DM.HR depends on the accuracy of extracted tuples, which is affected by all preprocessing steps. We computed the performance of tuple extraction by evaluating a sample of tuples extracted from a parsed version of SETIMES.HR against the tuples extracted from the SETIMES.HR gold annotations (we use the same sample as for tagging and parsing performance evaluation). Table 2 shows Precision, Recall, and F_1 score. Overall, we achieve the best performance on the *Atr* links, followed by *Pred* links. The performance is generally higher on unlexicalized links than on lexicalized links (note that performance on unlexical-

Link	Word	LMI	Link	Word	LMI
Atv	*moći*	225107	*Adv*	*moguće*	9669
Atv	*željeti*	22049	*Atv*	*namjeravati*	9095
Obj	*stan*	19997	*Obj*	*karta*	8936
po	*cijena*	18534	***prije***	*godina*	8584
Pred	*kada*	14408	*Adv*	*nedavno*	7842
Obj	*dionica*	13720	*Atv*	*odlučiti*	7578
Atv	*morati*	12097	*Adv*	*godina*	7496
Obj	*ulaznica*	11126	*Obj*	*zemljište*	7180

Table 3: Top 16 LMI-scored tuples for the verb *kupiti (to buy)*

ized *Verb* links is identical to overall performance on lexicalized verb links). The overall F_1 score of tuple extraction is 74.6%.

Following DM and DM.DE, we score each extracted tuple using Local Mutual Information (LMI) (Evert, 2005):

$$\mathrm{LMI}(i,j,k) = f(i,j,k) \log \frac{P(i,j,k)}{P(i)P(j)P(k)}$$

For a tuple (w_1, l, w_2), LMI scores the association strength between word w_1 and word w_2 via link l by comparing their joint distribution against the distribution under the independence assumption, multiplied with the observed frequency $f(w_1, l, w_2)$ to discount infrequent tuples. The probabilities are computed from tuple counts as maximum likelihood estimates. We exclude from the tensor all tuples with a negative LMI score. Finally, we symmetrize the tensor by introducing inverse links.

Model statistics. The resulting DM.HR tensor consists of 2.3M lemmas, 121M links and 165K link types (including inverse links). On average, each lemma has 53 links. This makes DM.HR more sparse than English DM (796 link types), but less sparse than German DM (220K link types; 22 links per lemma). Table 3 shows an example of the extracted tuples for the verb *kupiti (to buy)*. DM.HR tensor is freely available for download.[4]

5 Evaluating DM.HR

Task. We present a pilot evaluation DM.HR on a standard task from distributional semantics, namely synonym choice. In contrast to tasks like predicting word similarity We use the dataset created by Karan et al. (2012), with more than 11,000 synonym choice questions. Each question consists of one target word (nouns, verbs, and adjectives) with four synonym candidates (one is correct). The questions were extracted automatically from a machine-readable dictionary of Croatian. An example item is *težak (farmer): poljoprivrednik (farmer), umjetnost (art), radijacija (radiation), bod (point)*. We sampled from the dataset questions for nouns, verbs, and adjectives, with 1000 questions each.[5] Additionally, we manually corrected some errors in the dataset, introduced by the automatic extraction procedure. To make predictions, we compute pairwise cosine similarities of the target word vectors with the four candidates and predict the candidate(s) with maximal similarity (note that there may be ties).

	Accuracy (%)			Coverage (%)		
Model	N	A	V	N	A	V
DM.HR	**70.0**	66.3	**63.2**	99.9	99.1	100
BOW-LSA	67.2	**68.9**	61.0	100	100	100
BOW baseline	59.9	65.7	55.9	99.9	99.7	100

Table 4: Results on synonym choice task

Evaluation. Our evaluation follows the scheme developed by Mohammad et al. (2007), who define accuracy as the average number of correct predictions per covered question. Each correct prediction with a single most similar candidate receives a full credit (A), while ties for maximal similarity are discounted (B: two-way tie, C: three-way tie, D: four-way tie): $A + \frac{1}{2}B + \frac{1}{3}C + \frac{1}{4}D$. We consider a question item to be covered if the target and at least one answer word are modeled. In our experiments, ties occur when vector similarities are zero for all word pairs (due to vector sparsity). Note that a random baseline would perform at 0.25 accuracy.

As baseline to compare against the DM.HR, we build a standard bag-of-word model from the same corpus. It uses a ±5-word within-sentence context window, and the 10,000 most frequent context words (nouns, adjectives, and verbs) as dimensions. We also compare against BOW-LSA, a state-of-the-art synonym detection model from Karan et al. (2012), which uses 500 latent dimensions and paragraphs as contexts. We determine the significance of differences between the models by computing 95% confidence intervals with bootstrap resampling (Efron and Tibshirani, 1993).

Results. Table 4 shows the results for the three considered models on nouns (N), adjectives (A),

[4] http://takelab.fer.hr/dmhr

[5] Available at: http://takelab.fer.hr/crosyn

and verbs (V). The performance of BOW-LSA differs slightly from that reported by Karan et al. (2012), because we evaluate on a sample of their dataset. DM.HR outperforms the baseline BOW model for nouns and verbs (differences are significant at $p < 0.05$). Moreover, on these categories DM.HR performs slightly better than BOW-LSA, but the differences are not statistically significant. Conversely, on adjectives BOW-LSA performs slightly better than DM.HR, but the difference is again not statistically significant. All models achieve comparable and almost perfect coverage on this dataset (BOW-LSA achieves complete coverage because of the way how the original dataset was filtered).

Overall, the biggest improvement over the baseline is achieved for nouns. Nouns occur as heads and dependents of many link types (unlexicalized and lexicalized), and are thus well represented in the semantic space. On the other hand, adjectives seem to be less well modeled. Although the majority of adjectives occur as heads or dependents of the *Atr* relation, for which extraction accuracy is the highest (cf. Table 2), it is likely that a single link type is not sufficient. As noted by a reviewer, more insight could perhaps be gained by comparing the predictions of BOW-LSA and DM.HR models. The generally low performance on verbs suggests that their semantic is not fully covered in word- and syntax-based spaces.

6 Conclusion

We have described the construction of DM.HR, a syntax-based distributional memory for Croatian built from a dependency-parsed web corpus. To the best of our knowledge, DM.HR is the first freely available distributional memory for a Slavic language. We have conducted a preliminary evaluation of DM.HR on a synonym choice task, where DM.HR outperformed the bag-of-word model and performed comparable to an LSA model.

This work provides a starting point for a systematic study of dependency-based distributional semantics for Croatian and similar languages. Our first priority will be to analyze how corpus preprocessing and the choice of link types relates to model performance on different semantic tasks. Better modeling of adjectives and verbs is also an important topic for future research.

Acknowledgments

The first author was supported by the Croatian Science Foundation (project 02.03/162: "Derivational Semantic Models for Information Retrieval"). We thank the reviewers for their constructive comments. Special thanks to Hiko Schamoni, Tae-Gil Noh, and Mladen Karan for their assistance.

References

Željko Agić and Danijela Merkler. 2013. Three syntactic formalisms for data-driven dependency parsing of Croatian. *Proceedings of TSD 2013, Lecture Notes in Artificial Intelligence*.

Željko Agić, Marko Tadić, and Zdravko Dovedan. 2008. Improving part-of-speech tagging accuracy for Croatian by morphological analysis. *Informatica*, 32(4):445–451.

Željko Agić, Marko Tadić, and Zdravko Dovedan. 2009. Evaluating full lemmatization of Croatian texts. In *Recent Advances in Intelligent Information Systems*, pages 175–184. EXIT Warsaw.

Željko Agić. 2012. K-best spanning tree dependency parsing with verb valency lexicon reranking. In *Proceedings of COLING 2012: Posters*, pages 1–12, Bombay, India.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Bartosz Broda and Maciej Piasecki. 2008. Supermatrix: a general tool for lexical semantic knowledge acquisition. In *Speech and Language Technology*, volume 11, pages 239–254. Polish Phonetics Association.

Bartosz Broda, Magdalena Derwojedowa, Maciej Piasecki, and Stanisław Szpakowicz. 2008. Corpus-based semantic relatedness for the construction of Polish WordNet. In *Proceedings of LREC*, Marrakech, Morocco.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Tomaž Erjavec. 2012. MULTEXT-East: Morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46(1):131–142.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A Flexible, Corpus-driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics*, 36(4):723–763.

Stefan Evert. 2005. *The statistics of word cooccurrences*. Ph.D. thesis, PhD Dissertation, Stuttgart University.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos: An open source trigram tagger. In *Proceedings of ACL 2007*, pages 209–212, Prague, Czech Republic.

Zelig S. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Anton Karl Ingason, Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. 2008. A mixed method lemmatization algorithm using a hierarchy of linguistic identities (HOLI). In *Proceedings of GoTAL*, pages 205–216.

Vedrana Janković, Jan Šnajder, and Bojana Dalbelo Bašić. 2011. Random indexing distributional semantic models for Croatian language. In *Proceedings of Text, Speech and Dialogue*, pages 411–418, Plzeň, Czech Republic.

Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Distributional semantics approach to detecting synonyms in Croatian language. In *Proceedings of the Language Technologies Conference, Information Society*, Ljubljana, Slovenia.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWac: Compiling web corpora for Croatian and Slovene. In *Proceedings of Text, Speech and Dialogue*, pages 395–402, Plzeň, Czech Republic.

Nikola Ljubešić, Damir Boras, Nikola Bakarić, and Jasmina Njavro. 2008. Comparing measures of semantic similarity. In *Proceedings of the ITI 2008 30th International Conference of Information Technology Interfaces*, Cavtat, Croatia.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of CoNLL-X*, pages 216–220, New York, NY.

Olga Mitrofanova, Anton Mukhin, Polina Panicheva, and Vyacheslav Savitsky. 2007. Automatic word clustering in Russian texts. In *Proceedings of Text, Speech and Dialogue*, pages 85–91, Plzeň, Czech Republic.

Saif Mohammad, Iryna Gurevych, Graeme Hirst, and Torsten Zesch. 2007. Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of EMNLP/CoNLL*, pages 571–580, Prague, Czech Republic.

Preslav Nakov. 2001a. Latent semantic analysis for Bulgarian literature. In *Proceedings of Spring Conference of Bulgarian Mathematicians Union*, Borovets, Bulgaria.

Preslav Nakov. 2001b. Latent semantic analysis for Russian literature investigation. In *Proceedings of the 120 years Bulgarian Naval Academy Conference*.

Sebastian Padó and Jason Utt. 2012. A distributional memory for German. In *Proceedings of the KONVENS 2012 workshop on lexical-semantic resources and applications*, pages 462–470, Vienna, Austria.

Maciej Piasecki. 2009. Automated extraction of lexical meanings from corpus: A case study of potentialities and limitations. In *Representing Semantics in Digital Lexicography. Innovative Solutions for Lexical Entry Content in Slavic Lexicography*, pages 32–43. Institute of Slavic Studies, Polish Academy of Sciences.

Pavel Smrž and Pavel Rychlý. 2001. Finding semantically related words in large corpora. In *Text, Speech and Dialogue*, pages 108–115. Springer.

Marko Tadić. 2005. The Croatian Lemmatization Server. *Southern Journal of Linguistics*, 29(1):206–217.

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