Building and Evaluating a Distributional Memory for Croatian

Jan Šnajder*, Sebastian Padó†, and Željko Agić‡

*University of Zagreb, Faculty of Electrical Engineering and Computing
†Heidelberg University, Institut für Computerlinguistik
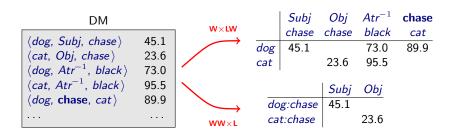‡University of Zagreb, Faculty of Humanities and Social Sciences

The 51st Annual Meeting of the
Association for Computational Linguistics
Sofia, August 7, 2013

Distributional semantics

- Representation of word meaning based on distributional hypothesis (Harris, 1954):
    - correlation between similarity of words' contexts and words' semantic similarity
    - words represented as vectors of context features
    - semantic similarity predicted via vector similarity
- Distributional semantic models used in many applications (Turney and Pantel, 2010)
- Most models use word-based or syntax-based co-occurrences
- Advantages of syntax-based models:
    - model fine-grained types of semantic similarity
    - capture long-distance contextual relationships
      $\Rightarrow$ important for free word order languages
    - applicable to various semantic tasks

Distributional memory (DM) (Baroni and Lenci, 2010)

- General, task-independent framework for distributional semantics
- Set of weighted Word-Link-Word triplets obtained from a corpus
  - links can be chosen to model dependency relations
- Task-specific sem. spaces obtained by arranging triplets into matrix

DM

⟨dog, Subj, chase⟩	45.1
⟨cat, Obj, chase⟩	23.6
⟨dog, Atr^{-1}, black⟩	73.0
⟨cat, Atr^{-1}, black⟩	95.5
⟨dog, chase, cat⟩	89.9
. . .	. . .

W×LW

	Subj chase	Obj chase	Atr^{-1} black	**chase** cat
dog	45.1		73.0	89.9
cat		23.6	95.5	

WW×L

	Subj	Obj
dog:chase	45.1	
cat:chase		23.6

- Dependency-based DM for English (Baroni and Lenci, 2010) and German (DM.DE) (Padó and Utt, 2012)

Building DM.HR

- Required:
  - good, clean, and large corpus
  - good linguistic preprocessing
- A challenge, because Croatian is an under-resourced and a morphologically complex language
- Steps in building DM.HR:
  1. Corpus preparation
  2. Tagging, lemmatization, and parsing
  3. Triplet extraction

Step 1: Corpus preparation

- Croatian web corpus hrWaC (Ljubešić and Erjavec, 2011)
- Boilerplate removed, but still contains non-parsable content
    - code snippets, encoding errors, non-diacriticized text, foreign-language content (Serbian, Slovenian, English, . . .)
- Additional heuristic filtering:
    1. website filter: blog/discussion forum content removed
    2. document filter: too short, foreign-language
    3. sentence filter: too short, non-standard symbols, non-diacriticized, foreign-language
- Filtered corpus fHrWaC: 51M sentences and 1.2G tokens

Step 2: Tagging, lemmatization, and parsing

- We trained the models on SETIMES.HR, the Croatian part of the SETimes parallel corpus
  - 90K tokens and 4K sentences
  - manually lemmatized and morphologically annotated
  - dependency annotated by Agić and Merkler (2013)
- HunPos tagger (Halácsy *et al.*, 2007)
- CST lemmatizer (Ingason *et al.*, 2008)
- MSTParser dependency parser (McDonald *et al.*, 2006)

Tagging, lemmatization, and parsing accuracy

		SETimes.Hr	Wikipedia
HunPos (POS only)	Acc	97.1	94.1
CST lemmatizer	Acc	97.7	96.5
MSTParser	LAS	77.5	68.8

- performance on Wikipedia: cross-domain evaluation
- state-of-the-art performance for Croatian
  - see (Agić and Merkler, 2013) and (Agić et al., 2013) for details

Step 3: Triplet extraction

- 10 unlexicalized link types:
    - main dependency relations: *Pred*, *Atr*, *Adv*, *Atv*, *Obj*, *Prep*, *Pnom*
    - subject subcategorization (*Sub_tr*/*Subj_intr*) to account for meaning shift due to verb reflexivization
      predati (to hand in): ⟨*student*, *Subj_tr*, *predati* ⟩
      predati se (to surrender): ⟨*trupe/troops*, *Subj_intr*, *predati* ⟩
    - an underspecified *Verb* link
- 2 lexicalized link types:
    - prepositions: ⟨*mjesto/place*, **na/on**, *sunce/sun* ⟩
    - verbs: ⟨*država/state*, **kupiti/buy**, *količina/amount* ⟩
- Triplets scored with local mutual information

$$\text{LMI}(w_1, l, w_2) = f(w_1, l, w_2) \log \frac{P(w_1, l, w_2)}{P(w_1)P(l)P(w_2)}$$

Triplet extraction accuracy

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

Dm.Hr

- 2.3M lemmas, 121M links and 165K link types
- top-scored (w_1, l, w_2) triplets for $w_1 = $ *kupiti (to buy)* :

l	w_2	LMI
Atv	*moći (can$_V$)*	225107
Atv	*željeti (wish$_V$)*	22049
Obj^{-1}	*stan (apartment$_N$)*	19997
po	*cijena (price$_N$)*	18534
Pred	*kada (when$_R$)*	14408
Obj^{-1}	*dionica (share$_N$)*	13720
Atv	*morati (must$_V$)*	12097
Obj^{-1}	*ulaznica (ticket$_N$)*	11126
Adv	*moguće (possible$_R$)*	9669
Atv	*namjeravati (intend$_V$)*	9095
Obj^{-1}	*karta (ticket$_N$)*	8936
...	...	...

Task-based evaluation

- Synonym choice – standard task from distributional semantics

Q: *težak (farmer)*

A:
  (a) *poljoprivrednik (agriculturist)*   (b) *umjetnost (art)*
  (c) *radijacija (radiation)*   (d) *bod (point)*

- Dataset: 1,000 question items for nouns, verbs, and adjectives, compiled from a machine readable dictionary (Karan *et al.*, 2012)

- Model: W×LW

- Prediction: Cosine similarity

- Evaluation: Accuracy (%) + Coverage (%)

Synonym choice: Results

	Accuracy (%)			Coverage (%)		
Model	N	A	V	N	A	V
Dm.Hr	**70.0**	66.3	**63.2**	99.9	99.1	100
LSA (Karan *et al.*, 2012)	67.2	**68.9**	61.0	100	100	100
BOW baseline	59.9	65.7	55.9	99.9	99.7	100

- Outperforms BOW and numerically outperforms LSA on N and V
- Differences across POSes
    - nouns: well modeled in syntactic space
    - adjectives: less well modeled (mostly occur with *Atr* links)
    - verbs: poorly modeled in word and syntactic spaces
- Nearly complete coverage

Summary

- DM.HR is a syntax-based DM for Croatian built from a dependency-parsed web corpus
  - first DM for a Slavic language
  - freely available from `takelab.fer.hr/dmhr`
- Evaluation on synonym choice task
  - DM.HR outperforms BOW, numerically outperforms LSA on N and V
- DM.HR can be used for a variety of semantic tasks
- Future work
  - better modeling of adjectives and verbs
  - influence of corpus preprocessing/link types

Acknowledgment

References I

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

Agić, v., Ljubešić, N., and Merkler, D. (2013). Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *Proceedings of BSNLP 2013*. In press.

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

Halácsy, P., Kornai, A., and Oravecz, C. (2007). HunPos: An open source trigram tagger. In *Proceedings of ACL 2007*, pages 209–212, Prague, Czech Republic.

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

Ingason, A. K., Helgadóttir, S., Loftsson, H., and Rögnvaldsson, E. (2008). A mixed method lemmatization algorithm using a hierarchy of linguistic identities (HOLI). In *Proceedings of GoTAL*, pages 205–216.

References II

Karan, M., Šnajder, J., and Dalbelo Bašić, B. (2012). Distributional semantics approach to detecting synonyms in Croatian language. In *Proceedings of the Language Technologies Conference, Information Society*, Ljubljana, Slovenia.

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

McDonald, R., Lerman, K., and Pereira, F. (2006). Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of CoNLL-X*, pages 216–220, New York, NY.

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

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