Thematic Roles and Semantic Space: Insights from Distributional Semantic Models

Gabriella Lapesa
University of Osnabrück
glapesa@uos.de

Stefan Evert
FAU Erlangen Nürnberg
stefan.evert@fau.de

The goal of this work is to use Distributional Semantic Models (Sahlgren, 2006; Turney, 2010) to get insights into the nature of thematic roles. In particular, we investigate whether the semantic representation produced by Distributional Semantic Models (henceforth, DSMs) is sensitive to effects of typicality involving thematic roles, and we quantify their relative prominence in the semantic representation encoded in the distributional space. Corpus-based modeling of selectional preferences and thematic fit is a well established field of research (see Erk et al., 2007 and references therein). What is peculiar to our approach is its attempt to model thematic fit data without taking into account syntactic relations, on the basis of distributional relatedness in bag-of-words DSMs. In this abstract we will show that (a) DSMs that make no use of syntax show good performances in a task related to selectional preference and (b) that the distribution of DSMs' performance across thematic relations shows patterns which are compatible with some general assumptions in theoretical linguistics.

1 Data and Models

Our computational simulation is based on experimental items from two studies which investigate event knowledge effects in semantic priming:

- Ferretti et al. (2001) showed that verbs facilitate the processing of nouns denoting prototypical participants in the depicted event (in the role of AGENT, PATIENT, INSTRUMENT), and of adjectives denoting features of prototypical participants (PATIENT FEATURE). The thematic role LOCATION did not show any priming effect.
- McRae et al. (2005) showed that nouns facilitate processing of verbs denoting events in which they are prototypical participants (in the role of AGENT, PATIENT, INSTRUMENT, LOCATION).

The set of stimuli from these two studies constitutes the gold standard of our evaluation task. Table 1 reports the number of triples for every thematic relation in the dataset and one example triple for each relation.

DSMs are evaluated in a classification task: given a target (e.g., *interview*) and the corresponding pair of primes in the dataset (*reporter* and *carpenter*, for the thematic role AGENT), we measure DSMs' accuracy in picking up the congruent prime on the basis of semantic distance. We expect the vectors for prototypical thematic role fillers to be closer to the respective verbs than the non-prototypical ones; in parallel we expect verbs to be closer to their prototypical fillers than to non-prototypical ones. Verbs and prototypical fillers co-occur, therefore, they occur in similar contexts. The reason why we expect verbs and prototypical fillers to be found closer in the semantic space is the presence of a shared a topic, namely, the event.

Our research consists of a large-scale evaluation of DSMs and their parameters. Evaluated parameters are:

Dataset	Relation	N	Prime$_c$	Prime$_i$	Target
Verb-Noun	AGENT	28	Pay	Govern	Customer
	PATIENT	18	Invite	Arrest	Guest
	PATIENT FEATURE	20	Comfort	Hire	Upset
	INSTRUMENT	26	Cut	Dust	Rag
	LOCATION	24	Confess	Dance	Court
Noun-Verb	AGENT	30	Reporter	Carpenter	Interview
	PATIENT	30	Bottle	Ball	Recycle
	INSTRUMENT	32	Chainsaw	Detergent	Cut
	LOCATION	24	Beach	Pub	Tan

Table 1: Experimental datasets from Ferretti et al. (2001) and McRae et al. (2005): number of pairs per relation and example stimuli

- Source corpus: BNC, WaCkypedia_EN, Wp500[1], UkWaC, and a combination of BNC, Wackypedia_EN, and UkWaC;
- Context window: 2, 5 and 15 words to the left and to the right of the target;
- Use of part-of-speech information: no part of speech information, part of speech information on the target, part of speech information on both targets and features;
- Scoring measure: frequency, simple log-likelihood, Mutual Information, t-score, z-score, Dice coefficient;
- Vector transformation: no transformation, logarithmic, sigmoid and root transformation;
- Distance measure: cosine, euclidean and manhattan distance;
- Dimensionality reduction: no dimensionality reduction, Random Indexing (1000 dimensions) and Randomized Singular Value decomposition (300 dimensions);
- Index of distributional relatedness: distance in the semantic space, backward rank (rank of prime in the neighbors of the target), forward rank (rank of target in the neighbors of the prime), average of backward and forward rank[2].

In total, 38,880 models were computed for all possible combinations of these parameters.

2 Results

Table 2 shows range and mean accuracy achieved by the DSMs for each thematic role. Results are reported for two indexes of distributional relatedness, namely *distance* and *forward rank* (position of the target in the ranked neighbors of the prime).

Dataset	Relation	Distance		Forward rank	
		Range	M	Range	M
Verb-Noun	AGENT	43-100	79.3	39-100	85.6
	PATIENT	44-100	83.4	50-100	87.8
	PATIENT FEATURE	35-95	72	40-100	81.2
	INSTRUMENT	42-100	80.2	38-100	82.6
	LOCATION	30-96	73.6	42-100	82.9
Noun-Verb	AGENT	40-100	77.1	47-100	87.5
	PATIENT	47-100	85.6	60-100	93.6
	INSTRUMENT	40-100	75.4	47-100	87.6
	LOCATION	42-96	79.4	46-96	85.2

Table 2: Identification of consistent prime on the basis of distributional relatedness

[1] A subset of WaCkypedia_EN, composed by the first 500 words of each article.

[2] The introduction and evaluation of this parameter has many implication for cognitive modeling, as rank can capture directionality in priming effects. We will not tackle this issue in this abstract for reasons of space.

First of all, the high performance achieved on all the relations shows that selectional preference is indeed a matter of topic. Even if we are not claiming that syntax does not play any role in selectional preference, the fact that such a high performance is achieved without using syntactic information suggests that verbs and their prototypical fillers can be interpreted as cues to event knowledge (for a review of this claim and of its consequences for lexical theories, see Elman, 2009).

A comparison between mean accuracies allows to rank thematic relations with respect to the robustness of the typicality effects shown by DSMs. The results of such ranking of thematic roles for the two datasets are:

- Ferretti et al. (2001)
  - Distance: PATIENT>INSTRUMENT>AGENT>LOCATION>PATIENT FEATURE
  - Forward rank: PATIENT>AGENT>LOCATION>INSTRUMENT>PATIENT FEATURE
- McRae et al. (2005)
  - Distance: PATIENT>LOCATION>AGENT>INSTRUMENT
  - Forward rank: PATIENT>INSTRUMENT>AGENT>LOCATION

These rankings reflect the relative saliency of thematic roles as event features in the DSM semantic space. They may also be interpreted in the light of distinctions commonly assumed in theoretical linguistics, namely between *arguments* and *adjuncts* and between the *internal* and *external* argument. In particular, the ranking PATIENT>AGENT>LOCATION> INSTRUMENT reported above may be mapped onto the scale of the syntactic proximity to the verb, the internal argument (e.g., THEME or PATIENT) being the closest to it, followed by the external argument (e.g., AGENT or CAUSE) and adjuncts (e.g., INSTRUMENT or LOCATION).

We display the best performing models for each relation in table 3 (index of distributional relatedness: forward rank). For each relation, we report the number of models that achieved the best accuracy and we specify one of the best models: this choice should not be considered representative of general trends of performance. Such trends need to be evaluated with a different type of analysis, given the high number of parameter combinations involved in our study.

Dataset	Relation	Best Model								
		Acc	N	Corpus	Win	Pos	Score	Trans	Dist	Dim.Red
V-N	AGENT	100	11	wacky	5	targ+feat	freq	none	man	ri
	PATIENT	100	825	wacky	15	target	MI	none	cos	ri
	PATIENT FEAT.	100	4	wacky	5	targ+feat	freq	none	cos	ri
	INSTRUMENT	100	22	joint	15	targ	freq	none	man	rsvd
	LOCATION	100	127	joint	15	no pos	t-sc	log	cos	none
N-V	AGENT	100	643	bnc	15	none	s-ll	none	euc	none
	PATIENT	100	2357	ukwac	5	targ+feat	freq	log	cos	none
	INSTRUMENT	100	302	ukwac	15	none	freq	none	cos	rsvd
	LOCATION	95.8	504	joint	15	none	s-ll	none	cos	none

Table 3: Identification of consistent prime on the basis of distributional relatedness, forward rank: best accuracy (*Acc*), number of models that achieved best accuracy (*N*), the set of parameters defining one of the best models

3 What we will present

This abstract sketches the general features of our study. In the presentation we will provide further details concerning the analysis of distribution of accuracy per thematic relation, and we will present an evaluation of the impact of the different parameters on the performance of the models.

References

Elman, Jeff L. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33(4), 547-582.

Ferretti, Todd, Ken McRae and Anne Hatherell. 2001. Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4), 516-547.

Erk, Katrin, Sebastian Padó and Ulrike Padó. 2007. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4), 723- 763.

McRae, Ken, Mary Hare, Jeff L. Elman and Todd Ferretti. (2005). A basis for generating expectancies for verbs from nouns. *Memory and Cognition*, 33(7), 1174-1184.

Sahlgren, Magnus. (2006). *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high- dimensional vector spaces*. Unpublished doctoral dissertation, University of Stockolm.

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