10.12871/CLICIT2014165

Combining unsupervised syntactic and semantic models of thematic fit

Asad Sayeed and Vera Demberg
Computational Linguistics and Phonetics / MMCI
Saarland University
D-66123 Saarbrücken
{asayeed,vera}@coli.uni-saarland.de

Abstract

English. We explore the use of the SENNA semantic role-labeller to define a distributional space to build a fully unsupervised model of event-entity thematic fit judgements. Existing models use syntactic dependencies for this. Our Distributional Memory model outperforms a syntax-based model by a wide margin, matches an augmented model that uses hand-crafted rules, and provides results that can be easily combined with the augmented model, improving matching over multiple thematic fit judgement tasks.

Italiano. *I giudizi di Thematic Fit tra eventi ed entità sono stati modellati in passato facendo ricorso a dipendenze sintattiche. Il nostro modello utilizza invece uno spazio distribuzionale costruito in maniera non supervisionata con un Semantic Role Labeler (SENNA). Il nostro modello ottiene risultati nettamente migliori rispetto a un modello basato su dipendenze sintattiche e comparabili a quelli di un modello potenziato, che sfrutta regole sviluppate manualmente in aggiunta alle dipendenze. Combinando il nostro modello e il modello potenziato si ottiene un ulteriore miglioramento dei risultati su diversi compiti di giudizio di Thematic Fit.*

1 Introduction

It is perfectly conceivable that automated tasks in natural language semantics can be accomplished entirely through models that do not require the contribution of semantic features to work at high accuracy. Unsupervised semantic role labellers such as that of Titov and Klementiev (2011) and Lang and Lapata (2011) do exactly this: predict semantic roles strictly from syntactic realizations. In other words, for practical purposes, the relevant and frequent semantic cases might be completely covered by learned syntactic information. For example, given a sentence *The newspaper was put on the table*, such SRL systems would identify that *the table* should receive a "location" role purely from the syntactic dependencies centered around the preposition *on*.

We could extend this thinking to a slightly different task: thematic fit modelling. It could well be the case that the *the table* could be judged a more appropriate filler of a location role for *put* than, e.g., *the perceptiveness*, entirely due to information about the frequency of word collocations and syntactic dependencies collected through corpus data, handmade grammars, and so on. In fact, today's distributional models used for modelling of selectional preference or thematic fit generally base their estimates on syntactic or string co-occurrence models (Baroni and Lenci, 2010; Ritter et al., 2010; Séaghdha, 2010). The Distributional Memory (DM) model by Baroni and Lenci (2010) is one example of an unsupervised model based on syntactic dependencies, which has been successfully applied to many different distributional similarity tasks, and also has been used in compositional models (Lenci, 2011).

While earlier work has shown that syntactic relations and thematic roles are related concepts (Levin, 1993), there are also a large number of cases where thematic roles assigned by a role labeller and their best-matching syntactic relations do not correspond (Palmer et al., 2005). However, it is possible that this non-correspondence is not a problem for estimating typical agents and patients from large amounts of data: agents will most of the time coincide with subjects, and patients will most of the time coincide with syntactic objects. On the other hand, the best resource

for estimating thematic fit should be based on labels that most closely correspond to the target task, i.e. semantic role labelling, instead of syntactic parsing. In this paper, we want to test how far a DM trained directly on a role labeller which produces PropBank style semantic annotations can complement the syntax-based DM model on thematic fit tasks, given a similar corpus of training data. We maintain the unsupervised nature of both models by combining their ratings by averaging without any weight estimation (we "guess" 50%) and show that we get an improvement in matching human judgements collected from previous experiments on agent/patient roles, location, and manner roles. We demonstrate that a fully unsupervised model based on a the SENNA role-labeller (Collobert et al., 2011) outperforms a corresponding model based on MaltParser dependencies (DepDM) by a wide margin. Furthermore, we show that the SENNA-based model can almost match B&L's better performing TypeDM model, which involves hand-crafted rules, and demonstrate that the SENNA-based model makes a contribution over and above the syntactic model in a range of thematic role labelling tasks.

1.1 Thematic role typicality

Thematic roles describe the relations that entities take in an event or relation. Thematic role fit correlates with human plausibility judgments (Padó et al., 2009; Vandekerckhove et al., 2009), which can be used to evaluate whether a distributional semantic model can be effectively encoded in the distributional space.

A suitable dataset is the plausibility judgment data set by Padó (2007), which includes 18 verbs with up to twelve nominal arguments, totalling 414 verb-noun-role triples. The words were chosen based on their frequency in the Penn Treebank and FrameNet. Human subjects were asked to how common the nominal arguments were as agents or as patients for the verbs. We also evaluate the DM models on a data set by McRae et al. (2005), which contains thematic role plausibility judgments for 1444 verb-role-noun triples calculated over the course of several experiments.

While the first two data sets only contain plausibility judgments for verbs and their agents and patients, we additionally use two data sets containing judgments for locations (274 verb-location pairs) and instruments (248 verb-instrument pairs) (McRae et al., 2005), to see how well these models apply to roles other than agent and patient. All ratings were on a scale of 1 to 7.

1.2 Semantic role labelling

Semantic role labelling (SRL) is the task of assigning semantic roles such as agent, patient, location, etc. to entities related to a verb or predicate. Structured lexica such as FrameNet, VerbNet and PropBank have been developed as resources which describe the roles a word can have and annotate them in text corpora such as the PTB. Both supervised and unsupervised techniques for SRL have been developed. Some build on top of a syntactic parser, while others work directly on word sequences. In this paper, we use SENNA, whose advantage is being very fast and robust (not needing parsed text) and is able to label large, noisy corpora such as UKWAC.

2 Distributional Memory

Baroni and Lenci (2010) present a framework for recording distributional information about linguistic co-occurrences in a manner explicitly designed to be multifunctional rather than being tightly designed to reflect a particular task. Distributional Memory (DM) takes the form of an order-3 tensor, where two of the tensor axes represent words or lemmas and the third axis represents the syntactic link between them.

B&L construct their tensor from a combination of corpora: the UKWAC corpus, consisting of crawled UK-based web pages, the British National Corpus (BNC), and a large amount of English Wikipedia. Their linking relation is based on the dependency-parser output of MaltParser (Nivre et al., 2007), where the links consist of lexicalized dependency paths and lexico-syntactic shallow patterns, selected by handcrafted rules.

The tensor is represented as a sparse array of triples of the form *(word, link, word)* with values as local mutual information (LMI), calculated as $O \log \frac{O}{E}$ where O is the observed occurrence count of the triple and E the count expected under independence. B&L propose different versions of representing the link between the words (encoding the link between the words in different degrees of detail) and ways of counting frequencies. Their DepDM model encodes the link as the (partially lexicalized) dependency path between words and counts occurrence frequencies of triples to cal-

model	*coverage* (%)	ρ
BagPack	100	60
ST-MeanDM	99	58
TypeDM	100	51
SENNA-DepDM	99	51
Padó	97	51
ParCos	98	48
DepDM	100	35

Table 1: Comparison on Padó data, results of other models from Baroni and Lenci (2010).

culate LMI. The more successful TypeDM model uses the same dependency path encoding as a link but bases the LMI estimates on type frequencies (counted over grammatical structures that link the words) rather than token frequencies.

The tensor also contains inverse links: if *(monster, sbj_tr eat)* appears in the tensor with a given LMI, another entry with the same LMI will appear as *(eat, sbj_tr*$^{-1}$*, monster)*.

B&L provide algorithms to perform computations relevant to various tasks in NLP and computational psycholinguistics. These operations are implemented by querying slices of the tensor. To assess the fit of a noun w_1 in a role r for a verb w_2, they construct a centroid from the 20 top fillers for r with w_2 selected by LMI, using subject and object link dependencies instead of thematic roles. To illustrate, in order to determine how well *table* fits as a location for *put*, they would construct a centroid of other locations for *put* that appear in the DM, e.g. *desk, shelf, account* ...

The cosine similarity between w_1's vector and the centroid represents the preference for the noun in that role for that verb. The centroid used to calculate the similarity represents the characteristics of the verb's typical role-fillers in all the other contexts in which they appear.

B&L test their procedure against the Padó et al. similarity judgements by using Spearman's ρ. They compare their model against the results of a series of other models, and find that they achieve full coverage of the data with a ρ of 0.51, higher than most of the other models except for the BagPack algorithm (Herdağdelen and Baroni, 2009), the only supervised system in the comparison, which achieved 0.60. Using the TypeDM tensor they freely provide, we replicated their result using our own tensor-processing implementation.

3 SENNA

SENNA (Collobert and Weston, 2007; Collobert et al., 2011) is a high performance role labeller well-suited for labelling a corpus the size of UKWAC and BNC due to its speed. It uses a multi-layer neural network architecture that learns in a sliding window over token sequences in a process similar to a conditional random field, working on raw text instead of syntactic parses. SENNA extracts features related to word identity, capitalization, and the last two characters of each word. From these features, the network derives features related to verb position, POS tags and chunking. It uses hidden layers to learn latent features from the texts which are relevant for the labelling task.

SENNA was trained on PropBank and large amounts of unlabelled data. It achieves a role labelling F score of 75.49%, which is slightly lower than state-of-the-art SRL systems which use parse trees as input (around 78% F score).

4 Implementation

We constructed a DM from the corpora used by B&L by running the sentences individually through SENNA and counting the *(assignee, role, assigner)* triples that emerged from the SENNA labelling. However, we omit the Wikipedia data included by Baroni and Lenci; results were better without them (ρ=48 on Padó), possibly an effect of genre.

SENNA assigns roles to entire phrases, but we only accepted head nouns and NN-composita. We used the part-of-speech tagging done by SENNA to identify head words and accepted only the first consecutive series of non-possessive noun-tagged words. If these are multiple words in this series (as in the case of composita), each of them is listed as a separate assignee. There is a very small amount of data loss due to parser errors and software crashes. Our implementation corresponds to B&L's DepDM model over MaltParser dependencies. The SENNA-based tensors are used to evaluate thematic fit data as in the method of B&L described above[1].

5 Experiments

We ran experiments with our tensor (henceforth SENNA-DepDM) on the following sources of thematic fit data: the Padó dataset, agents/patients from McRae, instrumental roles from McRae, and location roles from McRae. For each dataset, we calculated Spearman's ρ with respect to human plausibility judgments. We compared this against

[1]Our tensor will be provided via our web sites after this paper officially appears.

	TypeDM		SENNA-DepDM		ST-MeanDM		TDM/SENNA correl.
	cov. (%)	ρ	*cov.* (%)	ρ	*cov.* (%)	ρ	ρ
Padó	100	53	99	51	99	58	64
McRae agent/patient	95	32	96	24	95	32	59
McRae instrumental	93	36	94	19	92	38	23
McRae location	99	23	<100	19	<100	27	26

Table 2: Comparison of TypeDM to SENNA-DepDM and ST-MeanDM.

the performance of TypeDM given our implementation of B&L's thematic fit query system. We then took the average of the scores of SENNA-DepDM and TypeDM—we will call this ST-MeanDM—for each of these human judgement sources and likewise report ρ. We also report coverage for all these experiments.

During centroid construction, we used the ARG0 and ARG1 roles to find typical nouns for subject and object respectively. For the instrument role data, we mapped the verb-noun pairs to PropBank roles ARG2, ARG3 for verbs that have an INSTRUMENT in their frame, otherwise ARGM-MNR. We used "with" as the link for TypeDM-centroids; the same PropBank roles work with SENNA. For location roles, we used ARGM-LOC; TypeDM centroids are built with "in", "at", and "on" as locative prepositions.

6 Results and discussion

For all our results, we report coverage and Spearman's ρ. Spearman's ρ is calculated with missing items (due to absence in the tensor on which the result was based) removed from the calculation.

Our SENNA-based tensors are taken directly from SENNA output in a manner analogous to B&L's construction of DepDM from MaltParser dependency output. Both of them do much better than the reported results for DepDM (see Table 1) and one of them comes close to the performance of TypeDM on the Padó data. This suggests that improvements can be made to SENNA-DepDM by developing a procedure to determining lexicalized relation types mediated by PropBank roles, and calculating LMI values based on partially lexicalized types instead of tokens, similar to TypeDM.

Tables 2 shows that the MaltParser-based TypeDM and the SENNA-based DepDM models in combination achieve improved correlation with human judgments compared to TypeDM by itself[2]. The only exception was the McRae agent/patient data, which stayed the same. We also include the correlation between the TypeDM and SENNA-DepDM cosine similarities on each data set. These values suggest that even when their correlations with human judgements are similar, they only partly model the same aspects of thematic fit.

We calculated ρ on a per-verb basis for the Padó data on TypeDM and the SRL-augmented combined results and examined the differences. Augmentation by averaging with the SENNA-DepDM output improves ρ most strongly on verbs like "increase" and "ask". For example, SENNA-DepDM produces much sharper differences in judgements about whether "amount" can be the agent or patient of "increase", closer to human performance. Averaging with SENNA-DepDM also reduces the cosine similarities for both agent and patient roles of "state" with "ask", more in line with lower human judgements in both cases relative to the other nouns tested with "ask".

7 Conclusions

We have constructed a distributional memory based on SENNA-annotated thematic roles and shown an improved correlation with human data when combining it with the high-performing syntax-based TypeDM. We found that, even when built on similar corpora, SRL brings something to the table over and above syntactic parsing. In addition, our SENNA-based DM model was constructed in a manner roughly equivalent to B&L's simpler DepDM model, and yet it performs at a level far higher than DepDM on the Padó data set, on its own approaching the performance of TypeDM. It is likely that an SRL-based equivalent to TypeDM would further improve performance, and is thus a possible path for future work.

Our work also contributes the first evaluation of structured distributional models of semantics for thematic role plausibility for roles other than agent and patient.

[2]Baroni and Lenci used a version of the Pado data that erroneously swapped the judgments for some ARG0 vs. ARG1. We here evaluate on the original Pado data, with ARG2 for communicative verbs (*tell, ask, caution*) set to ARG1, as this is how SENNA labels the recipient of the utterances. This caused a small upward shift in the TypeDM results (from ρ=51 to 53), but should not cause DepDM (not made publicly available) to catch up.

References

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

Ronan Collobert and Jason Weston. 2007. Fast semantic extraction using a novel neural network architecture. In *Annual meeting-association for computational linguistics*, volume 45, page 560.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Amaç Herdağdelen and Marco Baroni. 2009. BagPack: A general framework to represent semantic relations. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 33–40, Athens, Greece, March. Association for Computational Linguistics.

Joel Lang and Mirella Lapata. 2011. Unsupervised semantic role induction via split-merge clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1117–1126, Portland, Oregon, USA, June. Association for Computational Linguistics.

Alessandro Lenci. 2011. Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2Nd Workshop on Cognitive Modeling and Computational Linguistics*, CMCL '11, pages 58–66, Stroudsburg, PA, USA. Association for Computational Linguistics.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

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

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Ulrike Padó, Matthew W. Crocker, and Frank Keller. 2009. A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, 33(5):794–838.

Ulrike Padó. 2007. *The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing*. Ph.D. thesis, Saarland University.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 424–434, Stroudsburg, PA, USA. Association for Computational Linguistics.

Diarmuid Ó. Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 435–444, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ivan Titov and Alexandre Klementiev. 2011. A bayesian model for unsupervised semantic parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1445–1455, Portland, Oregon, USA, June. Association for Computational Linguistics.

Bram Vandekerckhove, Dominiek Sandra, and Walter Daelemans. 2009. A robust and extensible exemplar-based model of thematic fit. In *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, pages 826–834.