Directional Distributional Similarity for Lexical Expansion

Lili Kotlerman, Ido Dagan, Idan Szpektor

Department of Computer Science Bar-Ilan University Ramat Gan, Israel lili.dav@gmail.com {dagan,szpekti}@cs.biu.ac.il Maayan Zhitomirsky-Geffet

Department of Information Science Bar-Ilan University Ramat Gan, Israel maayan.geffet@gmail.com

Abstract

Distributional word similarity is most commonly perceived as a symmetric relation. Yet, one of its major applications is lexical expansion, which is generally asymmetric. This paper investigates the nature of directional (asymmetric) similarity measures, which aim to quantify distributional feature inclusion. We identify desired properties of such measures, specify a particular one based on averaged precision, and demonstrate the empirical benefit of directional measures for expansion.

1 Introduction

Much work on automatic identification of semantically similar terms exploits Distributional Similarity, assuming that such terms appear in similar contexts. This has been now an active research area for a couple of decades (Hindle, 1990; Lin, 1998; Weeds and Weir, 2003).

This paper is motivated by one of the prominent applications of distributional similarity, namely identifying lexical expansions. Lexical expansion looks for terms whose meaning implies that of a given target term, such as a query. It is widely employed to overcome lexical variability in applications like Information Retrieval (IR), Information Extraction (IE) and Question Answering (QA). Often, distributional similarity measures are used to identify expanding terms (e.g. (Xu and Croft, 1996; Mandala et al., 1999)). Here we denote the relation between an expanding term u and an expanded term v as ' $u \rightarrow v'$.

While distributional similarity is most prominently modeled by symmetric measures, lexical expansion is in general a directional relation. In IR, for instance, a user looking for "baby food" will be satisfied with documents about "baby pap" or "baby juice" ('pap \rightarrow food', 'juice \rightarrow food'); but when looking for "frozen juice" she will not be satisfied by "frozen food". More generally, directional relations are abundant in NLP settings, making symmetric similarity measures less suitable for their identification.

Despite the need for directional similarity measures, their investigation counts, to the best of our knowledge, only few works (Weeds and Weir, 2003; Geffet and Dagan, 2005; Bhagat et al., 2007; Szpektor and Dagan, 2008; Michelbacher et al., 2007) and is utterly lacking. From an expansion perspective, the common expectation is that the context features characterizing an expanding word should be largely included in those of the expanded word.

This paper investigates the nature of directional similarity measures. We identify their desired properties, design a novel measure based on these properties, and demonstrate its empirical advantage in expansion settings over state-of-the-art measures¹. In broader prospect, we suggest that asymmetric measures might be more suitable than symmetric ones for many other settings as well.

2 Background

The distributional word similarity scheme follows two steps. First, a feature vector is constructed for each word by collecting context words as features. Each feature is assigned a weight indicating its "relevance" (or association) to the given word. Then, word vectors are compared by some vector similarity measure.

¹Our directional term-similarity resource will be available at http://aclweb.org/aclwiki/index.php? title=Textual_Entailment_Resource_Pool

To date, most distributional similarity research concentrated on symmetric measures, such as the widely cited and competitive (as shown in (Weeds and Weir, 2003)) *LIN* measure (Lin, 1998):

$$LIN(u, v) = \frac{\sum_{f \in FV_u \cap FV_v} [w_u(f) + w_v(f)]}{\sum_{f \in FV_u} w_u(f) + \sum_{f \in FV_v} w_v(f)}$$

where FV_x is the feature vector of a word x and $w_x(f)$ is the weight of the feature f in that word's vector, set to their pointwise mutual information.

Few works investigated a directional similarity approach. Weeds and Weir (2003) and Weeds et al. (2004) proposed a *precision* measure, denoted here *WeedsPrec*, for identifying the hyponymy relation and other generalization/specification cases. It quantifies the weighted coverage (or *inclusion*) of the candidate hyponym's features (u) by the hypernym's (v) features:

WeedsPrec
$$(u \to v) = \frac{\sum_{f \in FV_u \cap FV_v} w_u(f)}{\sum_{f \in FV_u} w_u(f)}$$

The assumption behind *WeedsPrec* is that if one word is indeed a generalization of the other then the features of the more specific word are likely to be included in those of the more general one (but not necessarily vice versa).

Extending this rationale to the textual entailment setting, Geffet and Dagan (2005) expected that if the meaning of a word u entails that of v then *all* its prominent context features (under a certain notion of "prominence") would be included in the feature vector of v as well. Their experiments indeed revealed a strong empirical correlation between such complete inclusion of prominent features and lexical entailment, based on web data. Yet, such complete inclusion cannot be feasibly assessed using an off-line corpus, due to the huge amount of required data.

Recently, (Szpektor and Dagan, 2008) tried identifying the entailment relation between lexical-syntactic templates using *WeedsPrec*, but observed that it tends to promote unreliable relations involving infrequent templates. To remedy this, they proposed to balance the directional *WeedsPrec* measure by multiplying it with the symmetric *LIN* measure, denoted here *balPrec*:

$$\textit{balPrec}(u \! \rightarrow \! v) \! = \! \sqrt{\textit{LIN}(u,v) \! \cdot \! \textit{WeedsPrec}(u \! \rightarrow \! v)}$$

Effectively, this measure penalizes infrequent templates having short feature vectors, as those usually yield low symmetric similarity with the longer vectors of more common templates.

3 A Statistical Inclusion Measure

Our research goal was to develop a directional similarity measure suitable for learning asymmetric relations, focusing empirically on lexical expansion. Thus, we aimed to quantify most effectively the above notion of *feature inclusion*.

For a candidate pair ' $u \rightarrow v$ ', we will refer to the set of *u*'s features, which are those tested for inclusion, as *tested features*. Amongst these features, those found in *v*'s feature vector are termed *included features*.

In preliminary data analysis of pairs of feature vectors, which correspond to a known set of valid and invalid expansions, we identified the following desired properties for a distributional inclusion measure. Such measure should reflect:

- 1. the proportion of included features amongst the tested ones (the core inclusion idea).
- 2. the relevance of included features to the expanding word.
- 3. the relevance of included features to the expanded word.
- 4. that inclusion detection is less reliable if the number of features of either expanding or expanded word is small.

3.1 Average Precision as the Basis for an Inclusion Measure

As our starting point we adapted the *Average Precision* (*AP*) metric, commonly used to score ranked lists such as query search results. This measure combines precision, relevance ranking and overall recall (Voorhees and Harman, 1999):

$$AP = \frac{\sum_{r=1}^{N} [P(r) \cdot rel(r)]}{\textit{total number of relevant documents}}$$

where r is the rank of a retrieved document amongst the N retrieved, rel(r) is an indicator function for the relevance of that document, and P(r) is precision at the given cut-off rank r.

In our case the feature vector of the expanded word is analogous to the set of all relevant documents while tested features correspond to retrieved documents. Included features thus correspond to relevant retrieved documents, yielding the following analogous measure in our terminology:

$$AP(u \to v) = \frac{\sum_{r=1}^{|FV_u|} [P(r) \cdot rel(f_r)]}{|FV_v|}$$
$$rel(f) = \begin{cases} 1, & \text{if } f \in FV_v \\ 0, & \text{if } f \notin FV_v \end{cases}$$
$$P(r) = \frac{|\text{included features in ranks 1 to } r|}{r}$$

where f_r is the feature at rank r in FV_u .

This analogy yields a feature inclusion measure that partly addresses the above desired properties. Its score increases with a larger number of included features (correlating with the 1^{st} property), while giving higher weight to highly ranked features of the expanding word (2^{nd} property).

To better meet the desired properties we introduce two modifications to the above measure. First, we use the number of tested features $|FV_u|$ for normalization instead of $|FV_v|$. This captures better the notion of feature inclusion (1st property), which targets the proportion of included features relative to the *tested* ones.

Second, in the classical AP formula all relevant documents are considered relevant to the same extent. However, features of the expanded word differ in their relevance within its vector (3rd property). We thus reformulate rel(f) to give higher relevance to highly ranked features in $|FV_v|$:

$$rel'(f) = \begin{cases} 1 - \frac{rank(f, FV_v)}{|FV_v| + 1} & \text{, if } f \in FV_v \\ 0 & \text{, if } f \notin FV_v \end{cases}$$

where $rank(f, FV_v)$ is the rank of f in FV_v .

Incorporating these two modifications yields the *APinc* measure:

$$APinc(u \rightarrow v) = \frac{\sum_{r=1}^{|FV_u|} [P(r) \cdot rel'(f_r)]}{|FV_u|}$$

Finally, we adopt the balancing approach in (Szpektor and Dagan, 2008), which, as explained in Section 2, penalizes similarity for infrequent words having fewer features (4th property) (in our version, we truncated *LIN* similarity lists after top 1000 words). This yields our proposed directional measure *balAPinc*:

$$\textit{balAPinc}(u \rightarrow v) = \sqrt{\textit{LIN}(u, v) \cdot \textit{APinc}(u \rightarrow v)}$$

4 Evaluation and Results

4.1 Evaluation Setting

We tested our similarity measure by evaluating its utility for lexical expansion, compared with baselines of the *LIN*, *WeedsPrec* and *balPrec* measures (Section 2) and a balanced version of *AP* (Section 3), denoted *balAP*. Feature vectors were created by parsing the Reuters RCV1 corpus and taking the words related to each term through a dependency relation as its features (coupled with the relation name and direction, as in (Lin, 1998)). We considered for expansion only terms that occur at least 10 times in the corpus, and as features only terms that occur at least twice.

As a typical lexical expansion task we used the ACE 2005 events dataset². This standard IE dataset specifies 33 event types, such as *Attack*, *Divorce*, and *Law Suit*, with all event mentions annotated in the corpus. For our lexical expansion evaluation we considered the first IE subtask of finding sentences that mention the event.

For each event we specified a set of representative words (*seeds*), by selecting typical terms for the event (4 on average) from its ACE definition. Next, for each similarity measure, the terms found similar to any of the event's seeds (' $u \rightarrow seed$ ') were taken as expansion terms. Finally, to measure the sole contribution of expansion, we removed from the corpus all sentences that contain a seed word and then extracted all sentences that contain expansion terms as mentioning the event. Each of these sentences was scored by the sum of similarity scores of its expansion terms.

To evaluate expansion quality we compared the ranked list of sentences for each event to the goldstandard annotation of event mentions, using the standard Average Precision (AP) evaluation measure. We report Mean Average Precision (MAP) for all events whose AP value is at least 0.1 for at least one of the tested measures³.

4.1.1 Results

Table 1 presents the results for the different tested measures over the ACE experiment. It shows that the symmetric *LIN* measure performs significantly worse than the directional measures, assessing that a directional approach is more suitable for the expansion task. In addition, balanced measures consistently perform better than unbalanced ones.

According to the results, *balAPinc* is the bestperforming measure. Its improvement over all other measures is statistically significant according to the two-sided Wilcoxon signed-rank test

²http://projects.ldc.upenn.edu/ace/, training part.

³The remaining events seemed useless for our comparative evaluation, since suitable expansion lists could not be found for them by any of the distributional methods.

LIN	WeedsPrec	balPrec	AP	balAP	balAPinc
0.068	0.044	0.237	0.089	0.202	0.312

Table 1: MAP scores of the tested measures on the ACE experiment.

seed	LIN	balAPinc	
death	murder, killing, inci-	suicide, killing, fatal-	
	dent, arrest, violence	ity, murder, mortality	
marry	divorce, murder, love,	divorce, remarry,	
	dress, abduct	father, kiss, care for	
arrest	detain, sentence,	detain, extradite,	
	charge, jail, convict	round up, apprehend,	
		imprison	
birth	abortion, pregnancy,	wedding day,	
	resumption, seizure,	dilation, birthdate,	
	passage	circumcision, triplet	
injure	wound, kill, shoot,	wound, maim, beat	
	detain, burn	up, stab, gun down	

Table 2: Top 5 expansion terms learned by *LIN* and *balAPinc* for a sample of ACE seed words.

(Wilcoxon, 1945) at the 0.01 level. Table 2 presents a sample of the top expansion terms learned for some ACE seeds with either *LIN* or *balAPinc*, demonstrating the more accurate expansions generated by *balAPinc*. These results support the design of our measure, based on the desired properties that emerged from preliminary data analysis for lexical expansion.

Finally, we note that in related experiments we observed statistically significant advantages of the *balAPinc* measure for an unsupervised text categorization task (on the 10 most frequent categories in the Reuters-21578 collection). In this setting, category names were taken as seeds and expanded by distributional similarity, further measuring cosine similarity with categorized documents similarly to IR query expansion. These experiments fall beyond the scope of this paper and will be included in a later and broader description of our work.

5 Conclusions and Future work

This paper advocates the use of directional similarity measures for lexical expansion, and potentially for other tasks, based on distributional inclusion of feature vectors. We first identified desired properties for an inclusion measure and accordingly designed a novel directional measure based on averaged precision. This measure yielded the best performance in our evaluations. More generally, the evaluations supported the advantage of multiple directional measures over the typical symmetric LIN measure.

Error analysis showed that many false sentence extractions were caused by ambiguous expanding and expanded words. In future work we plan to apply disambiguation techniques to address this problem. We also plan to evaluate the performance of directional measures in additional tasks, and compare it with additional symmetric measures.

Acknowledgements

This work was partially supported by the NEGEV project (www.negev-initiative.org), the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886 and by the Israel Science Foundation grant 1112/08.

References

- R. Bhagat, P. Pantel, and E. Hovy. 2007. LEDIR: An unsupervised algorithm for learning directionality of inference rules. In *Proceedings of EMNLP-CoNLL*.
- M. Geffet and I. Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of ACL*.
- D. Hindle. 1990. Noun classification from predicateargument structures. In *Proceedings of ACL*.
- D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*.
- R. Mandala, T. Tokunaga, and H. Tanaka. 1999. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of SI-GIR*.
- L. Michelbacher, S. Evert, and H. Schutze. 2007. Asymmetric association measures. In *Proceedings* of *RANLP*.
- I. Szpektor and I. Dagan. 2008. Learning entailment rules for unary templates. In *Proceedings of COL-ING*.
- E. M. Voorhees and D. K. Harman, editors. 1999. *The Seventh Text REtrieval Conference (TREC-7)*, volume 7. NIST.
- J. Weeds and D. Weir. 2003. A general framework for distributional similarity. In *Proceedings of EMNLP*.
- J. Weeds, D. Weir, and D. McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of COLING*.
- F. Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83.
- J. Xu and W. B. Croft. 1996. Query expansion using local and global document analysis. In *Proceedings* of *SIGIR*.