Distributional Lexical Entailment by Topic Coherence

Laura Rimell
University of Cambridge
Computer Laboratory
laura.rimell@cl.cam.ac.uk

Abstract

Automatic detection of lexical entailment, or hypernym detection, is an important NLP task. Recent hypernym detection measures have been based on the Distributional Inclusion Hypothesis (DIH). This paper assumes that the DIH sometimes fails, and investigates other ways of quantifying the relationship between the co-occurrence contexts of two terms. We consider the top features in a context vector as a topic, and introduce a new entailment detection measure based on Topic Coherence (TC). Our measure successfully detects hypernyms, and a TC-based family of measures contributes to multi-way relation classification.

1 Introduction

Automatically detecting lexical entailment – for example, that *lion* entails *animal* or *guitar* entails *instrument*, also known as hypernym detection – is an important linguistic task in its own right, and is also a prerequisite for recognizing entailments between longer text segments such as phrases or sentences (Bos and Markert, 2005; Garrette et al., 2011; Baroni et al., 2012; Beltagy et al., 2013).

Several recent techniques for hypernym detection have made use of distributional semantics (Weeds and Weir, 2003; Weeds et al., 2004; Clarke, 2009; Kotlerman et al., 2010; Lenci and Benotto, 2012). These techniques are based on the Distributional Inclusion Hypothesis (Geffet and Dagan, 2005), hereafter DIH, which proposes that if term A entails term B (B is a hypernym of A), then the contexts in which A occurs are a subset of those in which B occurs. For example, all the contexts (co-occurrences) of *lion* – which might include *zoo, hunt, wild, food*, etc. – are also contexts of *animal*. Existing measures look at the *amount* of overlap between the co-occurrences of A and B, in order to judge whether B is a hypernym of A.

The motivation for the present paper is the well-known fact that the DIH is not fully correct. There are many reasons why a hyponym might occur in contexts where its hypernym does not. Some contexts are collocational, e.g. *lion king*. Other contexts are highly specific, e.g. *mane* applies uniquely to lions, horses, and zebras; it would be unusual to see text about *animals* with *manes*. The need to be informative is also relevant: *lion cub* will occur much more frequently than *animal cub*, since *animal* is of the wrong level of generality to pair with *cub*.

Moreover, the more general a hypernym becomes – up to the level of WordNet root elements, such as *entity* – its predominant sense ceases to correspond to the sense intended in hyponym-hypernym chains. Thus we never hear about going to visit an *entity* at the *zoo*.

This paper starts from the assumption that the DIH sometimes fails, and investigates not the *amount of containment* of A's features in B's features, but rather the *nature of the non-contained features*. We consider the top features of a distributional vector as a *topic*, and use recent measures for automatically measuring Topic Coherence (Newman et al., 2010; Mimno et al., 2011) to evaluate how the topics change under various conditions. Using a notion of vector negation, we investigate whether the distributional topic of e.g. *lion* becomes more or less coherent when we subtract the contexts of *animal*.

We introduce a new measure, Ratio of Change in Topic Coherence (RCTC), for detecting lexical entailment. The measure detects hypernyms with reasonable accuracy, and a family of Topic Coherence measures is used to perform a multi-way classification of tuples by relation class. Finally, we investigate how the level of generality of a hypernym affects entailment measures.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 511–519,
Gothenburg, Sweden, April 26-30 2014.

2 Related Work

Historically, manually developed resources such as WordNet (Miller, 1995) have been used to supply lexical entailment information to NLP applications (Bos and Markert, 2005). More recently, a number of techniques for detecting lexical entailment have been developed using distributional semantics (Weeds and Weir, 2003; Weeds et al., 2004; Geffet and Dagan, 2005; Clarke, 2009; Kotlerman et al., 2010; Lenci and Benotto, 2012). These measures quantify to what extent the co-occurrence features of a term A are included in those of another term B, by a direct comparison of the distributional vectors $\vec{A}$ and $\vec{B}$. Kotlerman et al. (2010) use the notion of Average Precision from Information Retrieval to weight the relative importance of the overlapping features. Lenci and Benotto (2012) also check the extent to which B's features are not a subset of A's, as a proxy for the more general character of B. The success of these feature inclusion measures has provided general support for the DIH. Following Szpektor and Dagan (2008), inclusion measures are also sometimes balanced with similarity measures such as LIN similarity (Lin, 1998), to ensure that A and B are semantically related, since unrelated pairs that differ in frequency can mimic feature inclusion.

Previous distributional approaches to hypernym detection have generally involved a single measure, designed to rank hypernyms above other relation classes. Evaluation has largely involved either ranking or binary classification tasks, and there has been little work on using a variety of measures to distinguish multiple relation classes. Lenci and Benotto (2012) perform a ranking task using the multi-class BLESS dataset (Baroni and Lenci, 2011), but not a classification. We perform a multi-way classification using a variety of Topic Coherence measures. Recent Semantic Relation Classification shared tasks (SemEval-2010 Task 8, SemEval-2012 Task 2) are also relevant, though the relation classes and approaches have differed.

3 Topic Coherence for Distributional Lexical Entailment

The intuition behind our approach is to investigate whether term A, the candidate hyponym, has a coherent topic reflected in its distributional features, which apply only to A and not to its hypernym B. Consider A=*beer*, B=*beverage*. They may share features such as *drink*, *cold*, and *party*. But if we minimize or exclude B's features and examine the remaining features of A (we discuss how to do this in Section 3.3), we might be left with more specific features such as *pint*, *lager*, and *brew*.

If A and B share almost all contexts, we would be left with a set of uninformative features, merely corpus noise. If A and B share few contexts, there would be little change to A's topic when excluding B's features. Between the extremes, a range of change in A's topic is possible; we seek to quantify this change and relate it to entailment.

To do this we need a way of treating a distributional context vector as a topic. We treat the N highest-weighted context features in $\vec{A}$ as the topic of A (topicA). If we represent the vector $\vec{A} \equiv \{f_{c_i,A}\}_i$, where $f_{c_i,A}$ is the weighted co-occurrence value of context feature c_i, then topicA is a set $\{c_j\}, j \in 1...N$, of the N highest-weighted context features c_j in $\vec{A}$.

3.1 Hypotheses

We consider two opposing hypotheses.

Hypothesis 1: Removing hypernym B's features from topicA will **decrease** the coherence of topicA. If being a B is very important to being an A, then the collection of remaining features may become more random. Hypothesis 1 is consistent with the DIH, since it implies that the important features of A are also features of B.

As a corollary, removing A's features from B may not change the coherence of topicB very much. Since A is just an instance of B, topicB retains coherence (i.e. there's a lot to being an *animal* besides what's involved in being a *lion*).

Hypothesis 2: Removing hypernym B's features from topicA will **increase** the coherence of topicA. Perhaps A, by virtue of being more specific, occurs in a highly coherent set of contexts where B does not. Hypothesis 2 is inconsistent with the DIH, since it imples that a hyponym always has specific features which the hypernym does not share.

As a corollary, removing hyponym A's features from hypernym B might decrease the coherence of topicB, if removing specific features leaves only more general, less informative features behind.

3.2 Topic Coherence Measure

We use a Topic Coherence (TC) measure from recent work on automatic evalution of topics generated from corpora by latent variable models (Newman et al., 2010; Mimno et al., 2011; Stevens et

al., 2012). TC measures are applied to the top N words from a generated topic. They assign pairwise relatedness scores to the words, and return the mean or median from the word-pair scores.

We adopt the best method from Newman et al. (2010), equal to the median pairwise Pointwise Mutual Information (PMI) of the top N words, using Wikipedia as a background corpus for PMI.[1] The measure is given in Equation (1):

$$TC(\{c_j\}) = \text{median}(PMI(c_i, c_k), i, k \in 1...N, i < k) \quad (1)$$

where $\{c_j\}$ is the topic, and PMI is defined as:

$$PMI(c_i, c_k) = log\frac{p(c_i, c_k)}{p(c_i)p(c_k)} \quad (2)$$

We use intra-sentence co-occurrence in Wikipedia for calculating PMI.

Note that our definition of a topic, namely the top N features from a distributional vector, does not correspond to a topic generated by a latent variable model, because it does not have a probability distribution over words. However, the TC measures we adopt do not make use of such a probability distribution except for choosing the top N words from a topic, which are then treated as an unordered set for the pairwise operations. Newman et al. (2010) uses N=10, and Mimno et al. (2011) uses N=5...20; we investigate a range of N.

3.3 Vector Negation

For removing one topic from another, we draw on the concept of vector negation (Widdows and Peters, 2003; Widdows, 2003). Vector negation has proved useful for modeling word senses in Information Retrieval. For example, one might want to formulate a query for *suit* NOT *lawsuit*, which will retrieve terms such as *shirt* and *jacket* and exclude *plaintiff* and *damages*.

We test two versions of vector negation. The first, **Widdows** (Widdows, 2003), represents A NOT B as the projection of $\vec{A}$ onto $\vec{B}^{\perp}$, the subspace orthogonal to $\vec{B}$ in the vector space V. Specifically, $\vec{B}^{\perp} \equiv \{v \in V : v \cdot \vec{B} = 0\}$. The formula for Widdows A NOT B is:

$$\text{A NOT B} \equiv \vec{A} - \frac{\vec{A} \cdot \vec{B}}{|\vec{B}|^2}\vec{B} \quad (3)$$

The second, **Strict** negation, simply zeros out any context features of A that are non-zero in B:

$$f_{c_i,AnotB} \equiv \begin{cases} 0 & \text{if } f_{c_i,B} \neq 0 \\ f_{i,A} & \text{if } f_{c_i,B} = 0 \end{cases} \quad (4)$$

This measure is harsher than Widdows negation, which decreases the value of common features but does not remove them completely.

3.4 Generality Measure

Herbelot and Ganesalingam (2013) experiment with hypernym detection using a generality measure. They measure the Kullback-Leibler (KL) divergence (Eq. 5) between the probability distribution over context words for a term A, and the background probability distribution. The idea is that the greater the KL divergence, the more informative and therefore specific the term is, while hypernyms are likely to be more general.

$$D_{\text{KL}}(p(f_i|A)||p(f_i)) = \Sigma_i ln(\frac{p(f_i|A)}{p(f_i)})p(f_i) \quad (5)$$

Herbelot and Ganesalingam (2013) found that KL divergence on its own was not sufficient for successful hypernym detection. We experiment with it in combination with TC measures.

4 Methods

4.1 Context Vectors

We produced context vectors from a 2010 Wikipedia download, lemmatized using morpha (Minnen et al., 2001). The 10K most frequent lemmas in the corpus, minus common stop words and the 25 most frequent lemmas, served as the context features. Feature co-occurrences were counted in a 7-word window around the target lemma (three words each side of the target lemma), and limited to intra-sentence co-occurrences.

Co-occurrence counts were weighted using T-test. We chose T-test because it does not overemphasize infrequent features; however, early experiments with Positive PMI weighting showed the overall performance of our measures to be similar with both weighting schemes.

We benchmarked our context vectors on the WS353 word similarity task (Finkelstein et al., 2002) and found them to be of comparable accuracy with previous literature.

Rel Class	Target	Related Word	Total
HYPER	*alligator*	*animal*	638
COORD	*alligator*	*lizard*	1,760
MERO	*alligator*	*mouth*	1,402
RAND-N	*alligator*	*message*	3,253

Table 1: Examples from the BLESS subset; number of tuples per relation in the development set.

[1] In our case, Wikipedia is also the source corpus for our context vectors.

	Macroaverage				Microaverage			
	Relation Class				Relation Class			
Coherence of	**HYPER**	**MERO**	**COORD**	**RAND-N**	**HYPER**	**MERO**	**COORD**	**RAND-N**
TopicA	5.14 ±1.63	5.16 ±1.66	5.13 ±1.63	5.16 ±1.66	5.14 ±1.59	5.37 ±1.56	5.22 ±1.63	5.28 ±1.62
TopicAnotB	3.82 ±1.27	3.86 ±1.02	3.49 ±0.94	5.07 ±1.50	3.88 ±1.73	4.07 ±1.42	3.58 ±1.51	5.17 ±1.64
TopicA-TopicAnotB	1.32 ±1.54	1.30 ±1.28	1.64 ±1.58	0.09 ±0.43	1.26 ±1.86	1.30 ±1.49	1.64 ±1.92	0.11 ±0.83
TopicB	4.97 ±0.58	4.51 ±0.52	5.02 ±0.73	4.49 ±0.24	5.01 ±1.15	4.53 ±1.44	5.07 ±1.63	4.50 ±1.30
TopicBnotA	4.36 ±0.55	3.92 ±0.53	3.33 ±0.67	4.45 ±0.27	4.37 ±1.15	3.89 ±1.32	3.35 ±1.61	4.46 ±1.41
TopicB-TopicBnotA	0.61 ±0.69	0.59 ±0.48	1.68 ±0.88	0.04 ±0.14	0.64 ±1.34	0.64 ±1.33	1.72 ±2.07	0.04 ±0.77

Table 2: Average Topic Coherence measures on the development set, using N=10, Strict negation.

4.2 Evaluation Dataset

We used a subset of the BLESS dataset (Baroni and Lenci, 2011) as defined by Lenci and Benotto (2012). The entire dataset consists of 200 concrete nouns in 17 broad noun classes (e.g. clothing, amphibian/reptile, vegetable, container), participating in a variety of relations. The subset contains the relation classes hypernym (HYPER), coordinate (COORD, i.e. co-hyponym), meronym (MERO, i.e. part-of), and random-noun (RAND-N, an unrelated noun). It consists of 14,547 tuples in total. Table 1 gives an example of each relation class, along with the total number of tuples per class in the development data.

Since there was no pre-defined development-test split for the BLESS subset, we randomly selected half of the data for development. For each of the 17 broad noun classes, we randomly chose half of the target nouns, and included all their HYPER, COORD, MERO, and RAND-N tuples. This resulted in a development set consisting of 96 target nouns and 7,053 tuples; and a test set consisting of 104 nouns and 7,494 tuples.

5 Topic Coherence Behavior

We first investigate how topic coherence behaves across the four relation classes. Table 2 shows the average values and standard deviation of TC-related measures on the development data. The left-hand side gives macro-averages, where values are first averaged per-class for each target word, then averaged across the 96 target words in the development set. The right-hand side gives micro-averages across all tuples in the development set. The micro- and macro-averages are similar, and we report macro-averages from now on.[2]

Row 1 of Table 2 shows the original coherence of topicA, and row 2 the coherence of topicAnotB. Row 3 is simply the difference between the two, showing the absolute change in coherence. Rows 4-6 are analogous. In general, coherence values for A and B ranged from the 3's to the 6's, with very high coherence of 7 or 8 and very low coherence of 1 or 2. We did not normalize TC values.

Comparing rows 1 and 4, we see that the B topics are slightly less coherent than the A topics, probably due to the makeup of the dataset (B terms include hypernyms and random words, while A terms are concrete nouns).

Column 1 shows that removing hypernym B from A results in a decrease in coherence, from 5.14 to 3.82. The difference in coherence, 1.32 in this case, is shown in row 3. Removing A from B also results in a coherence decrease, but a much smaller one: only a 0.61 average absolute decrease. Because the starting coherence values of A and B may be different, we focus on the amount of change in coherence when we perform the negation (rows 3 and 6), rather than the absolute coherence of the negated vectors (rows 2 and 5).

Interestingly, column 2 shows that the behaviour of meronyms is almost identical to hypernyms. This is surprising for two reasons: first, meronyms are intuitively more specific than their holonyms; and second, previous studies tended to conflate hypernyms with coordinates rather than meronyms (Lenci and Benotto, 2012).

Column 3, rows 3 and 6, show that coordinates behave differently from hypernyms and meronyms. Vector negation in both directions results in a similar loss of coherence (1.64 and 1.68), reflecting the fact that coordinates have a symmetrical relationship. The average change is also greater, although there is a wide variance. In column 4, the coherence differences for random nouns are again symmetrical, but in this case very small, since a randomly selected noun will not share many contexts with the target word.

We can also define a TC-based similarity mea-

[2]Lenci and Benotto (2012) also report macro-averages, but our figures are not comparable to theirs, which are based on a nearest-neighbor analysis.

	Relation Class			
Measure	**HYPER**	**MERO**	**COORD**	**RAND-N**
TC Meet	5.36	5.12	5.98	3.62
LIN	0.41	0.41	0.48	0.22
GenKLA	4.89	4.89	4.89	4.89
GenKLB	4.60	4.49	5.01	4.95
DiffGenKL	0.29	0.40	-0.12	-0.05

Table 3: Average similarity and generality measures on the dev. set, using N=10, Strict negation.

sure. We define $\vec{A}$ MEET $\vec{B}$ as the intersection of two vectors, where each feature value $f_{c_i,A \text{ MEET } B} \equiv \min(f_{c_i,A}, f_{c_i,B})$. Table 3 shows TC(A MEET B), with LIN similarity (Lin, 1998) between A and B for comparison. We expect that if A and B are similar, their common features will form a coherent topic. Indeed hypernyms and meronyms have high values, with coordinates slightly higher and random nouns much lower.

Table 3 also shows the KL divergence-based generality measure from Section 3.4. Term B is slightly more general (lower score) than term A for hypernyms and meronyms. This may suggest that meronyms are more general distributionally than their holonyms, e.g. *leg* is a holonym of *alligator*, but also associated with many other animals.

Table 4 shows the topics for *owl* and its hypernym *creature*. Using Strict negation to create *owl* NOT *creature* causes a number of contexts to be removed from *owl*: *sized, owl, burrow, hawk, typical, medium, eagle, large, nest*. Instead, more idiosyncratic contexts rise to the top, including *northern*, *mexican*, *grouping*, and *bar* (as in an owl's markings). These idiosyncratic contexts are not mutually informative and cause a sizeable decrease in TC.

On the other hand, removing *owl* from *creature* does not decrease the coherence nearly as much. The contexts that are promoted – *fantastic, bizarre, fairy* – are mutually consistent with the other *creature* contexts.

So far our results support Hypothesis 1: removing B from A decreases its coherence. However, we hypothesize that this may not be the case for hypernyms at all levels of generality. Considering the pair *owl-chordate*, there is no change from topicA to topicAnotB. But *chordate* loses a sizeable amount of coherence when *owl* is removed; the topic changes from *primitive, ancestral, ancestor, evolution, lineage, basal, earliest, fossil, non-, neural* (TC 6.62), to *earliest, non-, neural, affinity, probable, genome, suspected, universally, group, approximation* (TC 3.60).

6 Hypernym Detection Measures

Since we use the same dataset as Lenci and Benotto (2012), we report the invCL measure introduced in that paper, which outperformed the other measures reported there, including those of Weeds and Weir (2003), Weeds et al. (2004), and Clarke (2009). Let f_A be the weight of feature f in $\vec{A}$, and let F_A be the set of features with non-zero weights in $\vec{A}$. Then we have:

$$\text{CL}(A,B) = \frac{\Sigma_{f \in F_A \cap F_b} \min(f_A, f_B)}{\Sigma_{f \in F_A} f_A} \quad (6)$$

$$\text{invCL}(A,B) = \sqrt{\text{CL}(A,B) * (1 - \text{CL}(B,A))} \quad (7)$$

We also report the balAPinc measure of Kotlerman et al. (2010), which is not included in the Lenci and Benotto (2012) evaluation. This measure begins with APinc, in which the features of A are ranked by weight, highest to lowest:

$$\text{APinc}(A,B) = \frac{\Sigma_{r \in 1 \ldots |F_A|} P(r) * rel(f_r)}{|F_A|} \quad (8)$$

where $P(r)$ is the "precision" at rank r, that is, how many of B's features are included at rank r in the features of A; and $rel(f_r)$ is a relevance feature reflecting how important f_r is in B (see Kotlerman et al. (2010) for details). The balanced version balAPinc is:

$$\text{balAPinc}(A,B) = \sqrt{LIN(A,B) * \text{APinc}(A,B)} \quad (9)$$

owl **(5.19)**	*owl* **not** *creature* **(3.25)**	*creature* **(5.91)**	*creature* **not** *owl* **(5.09)**	*owl* **meet** *creature* **(4.14)**
barn	barn	mythical	mythical	small
sized	grey	-like	supernatural	large
owl	northern	strange	alien	burrow
burrow	mexican	supernatural	legendary	night
hawk	falcon	magical	fantastic	elf
typical	creek	alien	bizarre	little
medium	mountains	evil	aquatic	giant
eagle	grouping	legendary	dangerous	prey
large	bar	giant	vicious	hunt
nest	california	resemble	fairy	purple

Table 4: Topics from the development data with Topic Coherence values.

	Widdows			
N =	**5**	**10**	**15**	**20**
HYPER	1.00	1.00	1.00	1.00
MERO	0.99	1.00	1.00	1.00
COORD	1.02	1.00	1.00	1.01
RAND-N	1.00	1.00	1.00	1.00
	Strict			
N =	**5**	**10**	**15**	**20**
HYPER	1.64	**1.42**	1.23	1.19
MERO	1.91	1.23	1.24	1.20
COORD	1.36	1.15	1.10	1.16
RAND-N	1.08	1.03	1.03	1.02

Table 5: RCTC with varying N and neg type.

We introduce a new measure, Ratio of Change in Topic Coherence (RCTC). Based on Section 5, we expect that for hypernyms the change in coherence from A to AnotB is greater than the change from B to BnotA. However, we cannot simply use the ratio (A-AnotB)/(B-BnotA), because the very small changes in the RAND-N class result in very small denominators and unstable values. Instead, we consider two ratios: the magnitude of TC(A) compared to TC(AnotB), and the magnitude of TC(B) compared to TC(BnotA). We take the ratio of these figures:

$$\mathrm{RCTC}(A,B) = \frac{\frac{TC(\mathrm{topicA})}{TC(\mathrm{topicAnotB})}}{\frac{TC(\mathrm{topicB})}{TC(\mathrm{topicBnotA})}} \quad (10)$$

If topicA is much more coherent than AnotB, the numerator will be relatively large. If topicB is not much more coherent than topicBnotA, the denominator will be relatively small. Both of these factors encourage RCTC to be larger.[3]

We also balanced RCTC with three different factors: LIN similarity, a generality ratio, and TC(MeetAB). In each case we calculated the balanced value as $\sqrt{RCTC * factor}$.

7 Experiments and Discussion

We first look at the effect of N (topic size) and negation type on RCTC on the development data (Table 5). It is clear that RCTC distinguishes relation types using Strict but not Widdows negation. We believe this is because, as the "harsher" version of negation, it allows less-related features to rise to the top of the topic and reveal greater differences in topic coherence. N=10 was the only value that ranked hypernyms the highest; we use N=10 for the remaining experiments.

[3]Although TC values are PMI values, which can be negative, in practice the median pairwise PMI is almost never negative, because there tend to be more positive than negative values among the pairwise comparisons. Therefore, we have not accounted for sign in the ratio. We have handled as special cases the few instances where TC(topicAnotB) or TC(topicBnotA) takes the value of $-$infinity due to zero co-occurrences between many of the features.

	invCL	**bal APinc**	**RCTC**	**RCTC bal LIN**	**RCTC bal GEN**	**RCTC bal MEET**
HYPER	**0.41**	**0.23**	**1.37**	**0.72**	**1.09**	**2.62**
MERO	0.39	0.22	1.28	0.70	1.06	2.51
COORD	0.38	0.22	1.44	0.71	1.05	2.50
RAND-N	0.25	0.10	1.03	0.46	1.01	1.92

Table 6: Hypernym identification on full dataset: average value by relation.

We then proceed to hypernym identification on the full dataset (Table 6). All measures we tested assigned the highest average value to hypernyms (in bold) compared to the other relations.

7.1 Ranking Task

Lenci and Benotto (2012) introduced a ranking task for hypernym detection on the BLESS data, which we replicate here. In this task a measure is used to rank all tuples from the data. The accuracy of the ranking is assessed from the point of view of each relation class. The goal is for hypernyms to have the highest accuracy of all the classes.

We report the Information Retrieval (IR) measure Mean Average Precision (MAP) for each class, following Lenci and Benotto (2012). We also report Mean R-Precision (RPrec), equal to the precision at rank R where R is the number of elements in the class. None of the measures we evaluated achieves the highest result for hypernyms[4], though invCL consistently performs better for hypernyms than do the other measures (Table 7).

Both MAP and RPrec give more weight to correct rankings near the top of the list, as is suitable for IR applications. In the context of hypernym detection, they could test a system's ability to find one or two good-quality hypernyms quickly from a set of candidates. However, these measures are less appropriate for testing whether a system can, in general, rank hypernyms over other relations. Therefore, we also report Mean Area Under the ROC Curve, or Wilcoxon-Mann-Whitney statistic (AUC), which gives equal weight to correct rankings at the top and bottom of the list, and also compensates for unbalanced data. Table 7 shows that RCTCbalMEET performs identically to invCL on the AUC measure. This comparison suggests that invCL is better at placing hypernyms

[4]Lenci and Benotto (2012) report a different result, possibly due to the use of different context vectors.

		invCL	balAPinc	RCTC	RCTC balLIN	RCTC balGEN	RCTC balMEET
RPrec	Hyper	**0.30**	0.25	0.17	0.20	0.12	0.19
	Mero	0.32	0.29	0.30	0.31	0.21	0.32
	Coord	0.39	0.43	0.27	0.42	0.27	0.40
	Rand-N	0.18	0.19	0.38	0.16	0.42	0.18
AUC	Hyper	**0.18**	0.17	0.16	0.17	0.14	**0.18**
	Mero	0.31	0.31	0.27	0.31	0.24	0.31
	Coord	0.38	0.39	0.25	0.39	0.28	0.37
	Rand-N	0.13	0.13	0.32	0.12	0.34	0.15
MAP	Hyper	**0.35**	0.30	0.22	0.24	0.17	0.24
	Mero	0.37	0.35	0.35	0.36	0.27	0.37
	Coord	0.41	0.46	0.30	0.45	0.32	0.43
	Rand-N	0.32	0.32	0.43	0.31	0.46	0.33

Table 7: Ranking results. Bold indicates best result for hypernyms by evaluation measure.

at the top of the ranking, but over the whole dataset the two measures rank hypernyms above other tuples equally.

7.2 Classification Task

We performed a four-way classification of tuples by relation class. We used LIBSVM (Chang and Lin, 2011). As described in Section 4.2, the BLESS data is unbalanced, with hypernyms – our target class – making up only about 9% of the data. To address this imbalance, we used LIBSVM's option to increase the cost associated with the smaller classes during parameter tuning and training. We based the weights on the development data only (HYPER: 9% of the data, weight factor 10; MERO: 20% of the data, weight factor 5; COORD: 25% of the data, weight factor 4).

We used LIBSVM's default Radial Basis Function kernel. On the development data we performed 10-fold cross-validation. We used LIBSVM's grid.py utility for tuning the parameters C and γ separately for each fold. We also tuned and trained models on the development data and tested them on the test data.

We used four sets of features (Table 8): (1) invCL on its own; (2) TC features; (3) all features (invCL, TC, plus additional similarity and generality measures); and (4) all except TC features.

The results of classification on the development data are shown in Table 9, and on the test data in Table 10. Although we report overall accuracy, this is a poor measure of classificaton quality for unbalanced data. The tables therefore provide the Precision, Recall, and F-score by relation class.

The overall accuracy is respectable, although it can be seen that the hypernym class was the most difficult to predict, despite weighting the cost function. Hypernyms may be particularly difficult to isolate given their similarity to meronyms and intermediate status between coordinates and random nouns on some of the features.

Feature	Description
invCL	Lenci's invCL(A, B) (Eq. 7)
topicA	$TC(A)$
topicAnotB	$TC(B)$
diffTopicA	$TC(A) - TC(A \text{ NOT } B)$
ratioTopicsA	$TC(A \text{ NOT } B)/TC(A)$
topicB	$TC(B)$
topicBnotA	$TC(B \text{ NOT } A)$
diffTopicB	$TC(B) - TC(B \text{ NOT } A)$
ratioTopicsB	$TC(B \text{ NOT } A)/TC(B)$
topicMeetAB	$TC(A \text{ MEET } B)$
ratioTopics1	$TC(A \text{ NOT } B)/TC(B \text{ NOT } A)$
ratioTopics2	diffTopicA / diffTopicB
DiffTopics1	diffTopicA - diffTopicB
DiffTopics2	diffTopicA + diffTopicB
RCTC	RCTC(A, B) (Eq. 10)
RCTCbalMEET	RCTCbalMEET(A, B)
APinc	Kotlerman's APinc(A, B) (Eq. 8)
balAPinc	Kotlerman's balAPinc(A, B) (Eq. 9)
LIN	LIN similarity
genKLA	$D_{\mathrm{KL}}(p(f_i\|A)\|\|p(f_i))$ (Eq. 5)
genKLB	$D_{\mathrm{KL}}(p(f_i\|B)\|\|p(f_i))$ (Eq. 5)
diffGenKL	genKLA - genKLB
ratioGenKL	genKLA / genKLB
RCTCbalLIN	RCTCbalLIN(A, B)
RCTCbalGEN	RCTCbalGEN(A, B)
RCTCbalInvCL	RCTC(A, B) bal. with invCL(A, B)

Table 8: Features used in classification experiment. InvCL; TC features; additional features.

Importantly, while previous work has focused on single measures such as invCL, the classification task highlights a key aspect of the TC approach. Because we can measure the TC of several different vectors for any given tuple (original terms, negated topics, intersection, etc.) we can perform multi-way classification much more accurately than with the invCL measure alone. Moreover, the TC features make an important contribution to the multi-way classification over and above invCL and other previous similarity and generality

Feature Set	Acc	Class	P	R	F
invCL	39.2	Hyper	29.2	19.6	22.5
		Mero	25.5	51.7	34.0
		Coord	19.3	26.4	21.3
		Rand-N	73.5	44.9	55.6
TC Feats	56.7	Hyper	20.3	**41.4**	**27.1**
		Mero	36.5	48.4	41.4
		Coord	**66.5**	54.5	59.5
		Rand-N	**87.1**	64.7	74.2
All except TC	59.2	Hyper	28.7	19.7	22.9
		Mero	35.1	**56.2**	43.2
		Coord	58.2	54.5	56.2
		Rand-N	85.5	71.0	77.5
All	64.0	Hyper	**30.5**	24.4	26.7
		Mero	**44.9**	44.6	**44.6**
		Coord	60.3	**65.6**	**62.8**
		Rand-N	80.0	**79.6**	**79.7**

Table 9: Classification results on development data using 10-fold cross-validation.

Feature Set	Acc	Class	P	R	F
invCL	42.2	Hyper	31.1	19.3	23.8
		Mero	32.6	54.3	40.7
		Coord	23.1	29.3	25.8
		Rand-N	75.8	48.2	59.0
TC Feats	56.2	Hyper	20.0	**45.1**	27.7
		Mero	36.7	42.9	40.0
		Coord	**64.2**	56.5	60.1
		Rand-N	**88.6**	64.5	74.6
All except TC	60.6	Hyper	23.9	17.9	20.5
		Mero	38.1	**56.4**	**45.5**
		Coord	58.2	56.1	57.1
		Rand-N	86.5	73.8	79.6
All	63.1	Hyper	**33.9**	28.6	**31.0**
		Mero	**44.1**	36.9	40.2
		Coord	57.2	**64.3**	**60.6**
		Rand-N	78.2	**81.5**	**79.8**

Table 10: Classification results on test data using development data as training.

measures, with the set of all features yielding the highest overall accuracy.

Another interesting result is that classification with the TC features alone results in much higher recall (though lower precision) for hypernyms than any of the other feature sets, and on the development data (Table 9) results in the highest F-score for hypernyms.

8 Hypernym Depth

We performed a simple preliminary experiment to test the speculation that the interaction between topics depends on the level of generality of the hypernym. Using the WordNet::Similarity package (Pedersen et al., 2004), we divided the development data into bins according to the depth of the hypernym from the WordNet root node. Table 11 shows average values by hypernym depth.

D	Qty	diffA	diffB	RCTC	invCL	balAPinc
1	1	0.66	0.27	1.08	0.15	0.01
3	35	0.33	0.16	1.12	0.44	0.23
5	108	0.32	-0.65	1.32	0.33	0.16
6	41	1.21	0.24	1.50	0.44	0.21
7	160	1.45	0.64	1.34	0.44	0.27
8	136	1.30	0.90	1.25	0.35	0.19
9	71	1.37	1.09	1.26	0.41	0.23
10	51	1.90	2.10	2.08	0.41	0.24
11	15	1.85	1.50	1.23	0.48	0.31
12	13	2.08	1.45	1.24	0.28	0.17
13	3	2.49	0.97	1.67	0.27	0.12
14	4	2.02	1.97	1.05	0.27	0.09

Table 11: Average value by depth D of hypernym.

There is a striking result for diffA, i.e. TC(topicA) - TC(topicAnotB): the deeper the hypernym in the WordNet hierarchy, the greater the value. This suggests that more abstract hypernyms have less interaction with their hyponyms' topics. A similar, though less pronounced, effect is seen for diffB. However, the three measures RCTC, invCL, and balAPinc remain relatively stable as the hypernym depth changes. While this is somewhat reassuring, these averages clearly have not yet captured the difficulty which the DIH encounters in individual cases such as *owl–chordate*.

9 Conclusions

We have introduced a set of Topic Coherence measures, particularly the Ratio of Change in Topic Coherence, to identify hypernyms. These measures perform comparably to previous hypernym detection measures on many tasks, while providing a different view of the relationship between the distributional vectors of two terms, and contributing to a more accurate multi-way relation classification, especially higher recall for hypernyms.

The approach presented here provides a starting point for entailment measures that do not rely solely on the Distributional Inclusion Hypothesis. One issue with the current proposal is that it tests for a single coherent distributional topic, whereas multiple senses may be represented in a word's top context features. Future work will integrate Word Sense Disambiguation methods into the Topic Coherence based lexical entailment approach.

Acknowledgments

This work is supported by EPSRC grant EP/I037512/1. We gratefully acknowledge helpful discussion from Stephen Clark, Tamara Polajnar, Julie Weeds, Jeremy Reffin, David Weir, and the anonymous reviewers.

References

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the EMNLP workshop on GEMS: GEometrical Models of natural language Semantics*, pages 1–10, Edinburgh.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of EACL*, pages 23–32.

Islam Beltagy, Cuong Chau, Gemma Boleda, Dan Garrette, Katrin Erk, and Raymond Mooney. 2013. Montague meets markov: Deep semantics with probabilistic logical form. In *Proceedings of *SEM*, pages 11–21, Atlanta, Georgia.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of HLT-EMNLP*, pages 628–635, Vancouver.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Daoud Clarke. 2009. Context-theoretic semantics for natural language: an overview. In *Proceedings of the EACL workshop on GEMS: GEometrical Models of natural language Semantics*, pages 112–119, Athens.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20:116–131.

Dan Garrette, Katrin Erk, and Raymond Mooney. 2011. Integrating logical representations with probabilistic information using Markov Logic. In *Proceedings of IWCS*, Oxford, UK.

M. Geffet and I. Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of ACL*, Michigan.

Aurélie Herbelot and Mohan Ganesalingam. 2013. Measuring semantic content in distributional vectors. In *Proceedings of ACL*.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16:359–389.

Alessandro Lenci and Giuli Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of *SEM*, pages 75–79, Montreal.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of ICML*, Madison, Wisconson.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of EMNLP*, pages 262–272, Edinburgh.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of NAACL*, pages 100–108, Los Angeles, California.

Ted Pedersen, Siddarth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity - measuring the relatedness of concepts. In *Proceedings of NAACL (Demonstration System)*, pages 38–41, Boston, MA.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Butler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of EMNLP*, pages 952–961, Jeju Island, Korea.

I. Szpektor and I. Dagan. 2008. Learning entailment rules for unary templates. In *Proceedings of COLING*, Manchester, UK.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of EMNLP*, pages 81–88, Sapporo, Japan.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of COLING*, pages 1015–1021, Geneva.

Dominic Widdows and Stanley Peters. 2003. Word vectors and quantum logic. In *Proceedings of the Eight Mathematics of Language Conference*, Bloomington, Indiana.

Dominic Widdows. 2003. Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. In *Proceedings of ACL*, pages 136–143, Sapporo, Japan.