# Comparison of Similarity Models for the Relation Discovery Task

**Ben Hachey**
School of Informatics
University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW
bhachey@inf.ed.ac.uk

## Abstract

We present results on the relation discovery task, which addresses some of the shortcomings of supervised relation extraction by applying minimally supervised methods. We describe a detailed experimental design that compares various configurations of conceptual representations and similarity measures across six different subsets of the ACE relation extraction data. Previous work on relation discovery used a semantic space based on a term-by-document matrix. We find that representations based on term co-occurrence perform significantly better. We also observe further improvements when reducing the dimensionality of the term co-occurrence matrix using probabilistic topic models, though these are not significant.

## 1 Introduction

This paper describes work that aims to improve upon previous approaches to identifying relationships between named objects in text (e.g., people, organisations, locations). Figure 1 contains several example sentences from the ACE 2005 corpus that contain relations and Figure 2 summarises the relations occurring in these sentences. So, for example, sentence 1 contains an *employment* relation between Lebron James and Nike, sentence 2 contains a *sports-affiliation* relation between Stig Toefting and Bolton and sentence 4 contains a *business* relation between Martha Stewart (she) and the board of directors (of Martha Stewart Living Omnimedia).

1 As for that $90 million shoe contract with Nike, it may be a good deal for James.
2 Toefting transferred to Bolton in February 2002 from German club Hamburg.
3 Toyoda founded the automaker in 1937 ... .
4 In a statement, she says she's stepping aside in the best interest of the company, but she will stay on the board of directors.

Figure 1: Example sentences from ACE 2005.

| Sent | Entity$_1$ | Entity$_2$ | Relation |
|---|---|---|---|
| 1 | Lebron James | Nike | Employ |
| 2 | Stig Toefting | Bolton | Sports-Aff |
| 2 | Stig Toefting | Hamburg | Sports-Aff |
| 3 | Kiichiro Toyoda | Toyota Corp | Founder |
| 4 | Martha Stewart | board | Business |

Figure 2: Example entity pairs and relation types.

Possible applications include identifying companies taking part in mergers/acquisitions from business newswire, which could be inserted into a corporate intelligence database. In the biomedical domain, we may want to identify relationships between genes and proteins from biomedical publications, e.g. Hirschman et al. (2004), to help scientists keep up-to-date on the literature. Or, we may want to identify disease and treatment relations in publications and textbooks, which can be used to help formalise medical knowledge and assist general practitioners in diagnosis, treatment and prognosis (Rosario and Hearst, 2004).

Another application scenario involves building networks of relationships from text collections that indicate the important entities in a domain and can be used to visualise interactions. The networks could provide an alternative to searching when interacting with a document collection. This could prove beneficial, for example, in investigative journalism. It might also be used for social science research using techniques from social network analysis (Marsden and Lin, 1982). In previ-

*Proceedings of the Workshop on Linguistic Distances*, pages 25–34,
Sydney, July 2006. ©2006 Association for Computational Linguistics

ous work, relations have been used for automatic text summarisation as a conceptual representation of sentence content in a sentence extraction framework (Filatova and Hatzivassiloglou, 2004).

In the next section, we motivate and introduce the relation discovery task, which addresses some of the shortcomings of conventional approaches to relation extraction (i.e. supervised learning or rule engineering) by applying minimally supervised methods.[1] A critical part of the relation discovery task is grouping entity pairs by their relation type. This is a clustering task and requires a robust conceptual representation of relation semantics and a measure of similarity between relations. In previous work (Hasegawa et al., 2004; Chen et al., 2005), the conceptual representation has been limited to term-by-document (TxD) models of relation semantics. The current work introduces a term co-occurrence (TxT) representation for the relation discovery task and shows that it performs significantly better than the TxD representation. We also explore dimensionality reduction techniques, which show a further improvement.

Section 3 presents a parameterisation of similarity models for relation discovery. For the purposes of the current work, this consists of the semantic representation for terms (i.e. how a term's context is modelled), dimensionality reduction technique (e.g. singular value decomposition, latent Dirichlet allocation), and the measure used to compute similarity.

We also build on the evaluation paradigm for relation discovery with a detailed, controlled experimental setup. Section 4 describes the experiment design, which compares the various system configurations across six different subsets of the relation extraction data from the automatic content extraction (ACE) evaluation. Finally, Section 5 presents results and statistical analysis.

## 2 The Relation Discovery Task

Conventionally, relation extraction is considered to be part of information extraction and has been approached through supervised learning or rule engineering (e.g., Blaschke and Valencia (2002), Bunescu and Mooney (2005)). However, traditional approaches have several shortcomings. First and foremost, they are generally based on predefined templates of what types of relations exist in the data and thus only capture information whose importance was anticipated by the template designers. This poses reliability problems when predicting new data in the same domain as the training data will be from a certain epoch in the past. Due to language change and topical variation, as time passes, it is likely that the new data will deviate more and more from the trained models. Additionally, there are cost problems associated with the conventional supervised approach when updating templates or transferring to a new domain, both of which require substantial effort in re-engineering rules or re-annotating training data.

The goal of the relation discovery task is to identify the existence of associations between entities, to identify the kinds of relations that occur in a corpus and to annotate particular associations with relation types. These goals correspond to the three main steps in a generalised algorithm (Hasegawa et al., 2004):

1. Identify co-occurring pairs of named entities

2. Group entity pairs using the textual context

3. Label each cluster of entity pairs

The first step is the relation identification task. In the current work, this is assumed to have been done already. We use the gold standard relations in the ACE data in order to isolate the performance of the second step. The second step is a clustering task and as such it is necessary to compute similarity between the co-occurring pairs of named entities (relations). In order to do this, a model of relation similarity is required, which is the focus of the current work.

We also assume that it is possible to perform the third step.[2] The evaluation we present here looks just at the quality of the clustering and does not attempt to assess the labelling task.

## 3 Modelling Relation Similarity

The possible space of models for relation similarity can be explored in a principled manner by parameterisation. In this section, we discuss several

[1] The relation discovery task is minimally supervised in the sense that it relies on having certain resources such as named entity recognition. The focus of the current paper is the unsupervised task of clustering relations.

[2] Previous approaches select labels from the collection of context words for a relation cluster (Hasegawa et al., 2004; Zhang et al., 2005). Chen et al. (2005) use discriminative category matching to make sure that selected labels are also able to differentiate between clusters.

parameters including the term context representation, whether or not we apply dimensionality reduction, and what similarity measure we use.

### 3.1 Term Context

Representing texts in such a way that they can be compared is a familiar problem from the fields of information retrieval (IR), text mining (TM), textual data analysis (TDA) and natural language processing (NLP) (Lebart and Rajman, 2000). The traditional model for IR and TM is based on a term-by-document (TxD) vector representation. Previous approaches to relation discovery (Hasegawa et al., 2004; Chen et al., 2005) have been limited to TxD representations, using *tf*idf* weighting and the cosine similarity measure. In information retrieval, the weighted term representation works well as the comparison is generally between pieces of text with large context vectors. In the relation discovery task, though, the term contexts (as we will define them in Section 4) can be very small, often consisting of only one or two words. This means that a term-based similarity matrix between entity pairs is very sparse, which may pose problems for performing reliable clustering.

An alternative method widely used in NLP and cognitive science is to represent a term context by its neighbouring words as opposed to the documents in which it occurs. This term co-occurrence (TxT) model is based on the intuition that two words are semantically similar if they appear in a similar set of contexts (see e.g. Pado and Lapata (2003)). The current work explores such a term co-occurrence (TxT) representation based on the hypothesis that it will provide a more robust representation of relation contexts and help overcome the sparsity problems associated with weighted term representations in the relation discovery task. This is compared to a baseline term-by-document (TxD) representation which is a re-implementation of the approach used by Hasegawa et al. (2004) and Chen et al. (2005).

### 3.2 Dimensionality Reduction

Dimensionality reduction techniques for document and corpus modelling aim to reduce description length and model a type of semantic similarity that is more linguistic in nature (e.g., see Landauer et al.'s (1998) discussion of LSA and synonym tests). In the current work, we explore singular value decomposition (Berry et al., 1994), a technique from linear algebra that has been applied to a number of tasks from NLP and cognitive modelling. We also explore latent Dirichlet allocation, a probabilistic technique analogous to singular value decomposition whose contribution to NLP has not been as thoroughly explored.

Singular value decomposition (SVD) has been used extensively for the analysis of lexical semantics under the name of latent semantic analysis (Landauer et al., 1998). Here, a rectangular matrix is decomposed into the product of three matrices ($X_{w \times p} = W_{w \times n} S_{n \times n} (P_{p \times n})^T$) with $n$ 'latent semantic' dimensions. The resulting decomposition can be viewed as a rotation of the $n$-dimensional axes such that the first axis runs along the direction of largest variation among the documents (Manning and Schütze, 1999). $W$ and $P$ represent terms and documents in the new space. And $S$ is a diagonal matrix of singular values in decreasing order.

Taking the product $W_{w \times k} S_{k \times k} (P_{p \times k})^T$ over the first $D$ columns gives the best least square approximation of the original matrix $X$ by a matrix of rank $D$, i.e. a reduction of the original matrix to $D$ dimensions. SVD can equally be applied to the word co-occurrence matrices obtained in the TxT representation presented in Section 2, in which case we can think of the original matrix as being a term × co-occurring term feature matrix.

While SVD has proved successful and has been adapted for tasks such as word sense discrimination (Schütze, 1998), its behaviour is not easy to interpret. Probabilistic LSA (pLSA) is a generative probabilistic version of LSA (Hofmann, 2001). This models each word in a document as a sample from a mixture model, but does not provide a probabilistic model at the document level. Latent Dirichlet Allocation (LDA) addresses this by representing documents as random mixtures over latent topics (Blei et al., 2003). Besides having a clear probabilistic interpretation, an additional advantage of these models is that they have intuitive graphical representations.

Figure 3 contains a graphical representation of the LDA model as applied to TxT word co-occurrence matrices in standard plate notation. This models the word features $f$ in the co-occurrence context (size $N$) of each word $w$ (where $w \in \mathcal{W}$ and $|\mathcal{W}| = W$) with a mixture of topics $z$. In its generative mode, the LDA model samples a topic from the word-specific multino-

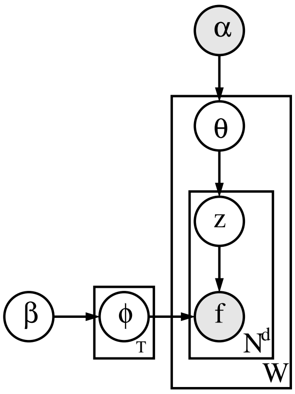

Figure 3: Graphical representation of LDA.

mial distribution $\theta$. Then, each context feature is generated by sampling from a topic-specific multinomial distribution $\phi_z$.[3] In a manner analogous to the SVD model, we use the distribution over topics for a word $w$ to represent its semantics and we use the average topic distribution over all context words to represent the conceptual content of an entity pair context.

### 3.3 Measuring Similarity

Cosine (Cos) is commonly used in the literature to compute similarities between *tf*idf* vectors:

$$Cos(p,q) = \frac{\sum_i p_i q_i}{\sqrt{\sum p^2}\sqrt{\sum q^2}}$$

In the current work, we use cosine over term and SVD representations of entity pair context. However, it is not clear which similarity measure should be used for the probabilistic topic models. Dagan et al. (1997) find that the symmetric information radius measure performs best on a pseudo-word sense disambiguation task, while Lee (1999) find that the asymmetric skew divergence – a generalisation of Kullback-Leibler divergence – performs best for improving probability estimates for unseen word co-occurrences.

In the current work, we compare KL divergence with two methods for deriving a symmetric measure. The KL divergence of two probability distributions ($p$ and $q$) over the same event space is defined as:

$$KL(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$$

In information-theoretic terms, KL divergence is the average number of bits wasted by encoding events from a distribution $p$ with a code based on distribution $q$. The symmetric measures are defined as:

$$Sym(p,q) = \frac{1}{2}\left[KL(p||q) + KL(q||p)\right]$$

$$JS(p,q) = \frac{1}{2}\left[KL\left(p||\frac{p+q}{2}\right) + KL\left(q||\frac{p+q}{2}\right)\right]$$

The first is termed symmetrised KL divergence (Sym) and the second is termed Jensen-Shannon (JS) divergence. We explore KL divergence as well as the symmetric measures as it is not known in advance whether a domain is symmetric or not.

Technically, the divergence measures are dissimilarity measures as they calculate the difference between two distributions. However, they can be converted to increasing measures of similarity through various transformations. We treated this as a parameter to be tuned during development and considered two approaches. The first is from Dagan et al. (1997). For KL divergence, this function is defined as $Sim(p,q) = 10^{-\beta KL(p||q)}$, where $\beta$ is a free parameter, which is tuned on the development set (as described in Section 4.2). The same procedure is applied for symmetric KL divergence and JS divergence. The second approach is from Lee (1999). Here similarity for KL is defined as $Sim(p,q) = C - KL(p||q)$, where $C$ is a free parameter to be tuned.

## 4 Experimental Setup

### 4.1 Materials

Following Chen et al. (2005), we derive our relation discovery data from the automatic content extraction (ACE) 2004 and 2005 materials for evaluation of information extraction.[4] This is preferable to using the New York Times data used by Hasegawa et al. (2004) as it has gold standard annotation, which can be used for unbiased evaluation.

The relation clustering data is based on the gold standard relations in the information extraction

[3] The hyperparameters $\alpha$ and $\beta$ are Dirichlet priors on the multinomial distributions for word features ($\phi \sim Dir(\beta)$) and topics ($\theta \sim Dir(\alpha)$). The choice of the Dirichlet is explained by its conjugacy to the multinomial distribution, meaning that if the parameter (e.g. $\phi$, $\theta$) for a multinomial distribution is endowed with a Dirichlet prior then the posterior will also be a Dirichlet. Intuitively, it is a distribution over distributions used to encode prior knowledge about the parameters ($\phi$ and $\theta$) of the multinomial distributions for word features and topics. Practically, it allows efficient estimation of the joint distribution over word features and topics $P(\vec{f}, \vec{z})$ by integrating out $\phi$ and $\theta$.

[4] `http://www.nist.gov/speech/tests/ace/`

data. We only consider data from newswire or broadcast news sources. We constructed six data subsets from the ACE corpus based on four of the ACE entities: persons (PER), organisations (ORG), geographical/social/political entities (GPE) and facilities (FAC). The six data subsets were chosen during development based on a lower limit of 50 for the data subset size (i.e. the number of entity pairs in the domain), ensuring that there is a reasonable amount of data. We also set a lower limit of 3 for the number of classes (relation types) in a data subset, ensuring that the clustering task is not too simple.

The entity pair instances for clustering were chosen based on several criteria. First, we do not use ACE's *discourse* relations, which are relations in which the entity referred to is not an official entity according to world knowledge. Second, we only use pairs with one or more non-stop words in the intervening context, that is the context between the two entity heads.[5] Finally, we only keep relation classes with 3 or more members. Table 4.1 contains the full list of relation types from the subsets of ACE that we used. (Refer to Table 4.2 for definition of the relation type abbreviations.)

We use the Infomap tool[6] for singular value decomposition of TxT matrices and compute the conceptual content of an entity pair context as the average over the reduced $D$-dimensional representation of the co-occurrence vector of the terms in the relation context. For LDA, we use Steyvers and Griffiths' Topic Modeling Toolbox[7]). The input is produced by a version of Infomap which was modified to output the TxT matrix. Again, we compute the conceptual content of an entity pair as the average over the topic vectors for the context words. As documents are explicitly modelled in the LDA model, we input a matrix with raw frequencies. In the TxD, unreduced TxT and SVD models we use *tf*idf* term weighting.

We use the same preprocessing when preparing the text for building the SVD and probabilistic topic models as we use for processing the intervening context of entity pairs. This consisted of Mx-Terminator (Reynar and Ratnaparkhi., 1997) for sentence boundary detection, the Penn Treebank sed script[8] for tokenisation, and the Infomap stop word list. We also use an implementation of the Porter algorithm (Porter, 1980) for stemming.[9]

### 4.2 Model Selection

We used the ACE 2004 relation data to perform model selection. Firstly, dimensionality ($D$) needs to be optimised for SVD and LDA. SVD was found to perform best with the number of dimensions set to $10$. For LDA, dimensionality interacts with the divergence-to-similarity conversion so they were tuned jointly. The optimal configuration varies by the divergence measure with $D = 50$ and $C = 14$ for KL divergence, $D = 200$ and $C = 4$ for symmetrised KL, and $D = 150$ and $C = 2$ for JS divergence. For all divergence measures, Lee's (1999) method outperformed Dagan et al.'s (1997) method. Also for all divergence measures, the model hyper-parameter $\beta$ was found to be optimal at $0.0001$. The $\alpha$ hyper-parameter was always set to $50/T$ following Griffiths and Steyvers (2004).

Clustering is performed with the CLUTO software[10] and the technique used is identical across models. Agglomerative clustering is used for comparability with the original relation discovery work of Hasegawa et al. (2004). This choice was motivated because as it is not known in advance how many clusters there should be in a new domain.

One way to view the clustering problem is as an optimisation process where an optimal clustering is chosen with respect to a criterion function over the entire solution. The criterion function used here was chosen based on performance on the development data. We compared a number of criterion functions including single link, complete link, group average, $\mathcal{I}_1$, $\mathcal{I}_2$, $\mathcal{E}_1$ and $\mathcal{H}_1$. $\mathcal{I}_1$ is a criterion function that maximises sum of pairwise similarities between relation instances assigned to each cluster, $\mathcal{I}_2$ is an internal criterion function that maximises the similarity between each relation instance and the centroid of the cluster it is assigned to, $\mathcal{E}_1$ is an external criterion function that minimises the similarity between the centroid vector of each cluster and the centroid vector of the

[5] Following results reported by Chen et al. (2005), who tried unsuccessfully to incorporate words from the surrounding context to represent a relation's semantics, we use only intervening words.

[6] http://infomap.stanford.edu/

[7] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

[8] http://www.cis.upenn.edu/˜treebank/tokenizer.sed

[9] http://www.ldc.usb.ve/˜vdaniel/porter.pm

[10] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

| ORG-GPE | | ORG-ORG | | PER-FAC | | PER-GPE | | PER-ORG | | PER-PER | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| basedin | 54 | subsidiary | 36 | located | 127 | located | 222 | staff | 121 | business | 81 |
| subsidiary | 27 | emporgothr | 14 | owner | 14 | resident | 79 | executive | 100 | family | 20 |
| located | 15 | partner | 8 | near | 4 | executive | 42 | member | 44 | persocothr | 16 |
| gpeaffothr | 3 | member | 6 | | | staff | 30 | emporgothr | 27 | perorgothr | 9 |
| | | | | | | employgen | 7 | employgen | 9 | near | 7 |
| | | | | | | | | located | 4 | ethnic | 5 |
| | | | | | | | | | | executive | 3 |
| | | | | | | | | | | ideology | 3 |
| | | | | | | | | | | member | 3 |
| *Total* | 99 | *Total* | 64 | *Total* | 145 | *Total* | 380 | *Total* | 305 | *Total* | 147 |

Table 1: Relation distributions for entity pair domains.

| Type | Subtype | Abbr |
|---|---|---|
| AGENT-ARTIFACT | User-or-Owner | owner |
| EMPLOY/MEMBER | Employ-Executive | executive |
| | Employ-Staff | staff |
| | Employ-Undet'd | employgen |
| | Member-of-Group | member |
| | Other | artothr |
| | Partner | partner |
| | Subsidiary | subsidiary |
| GPE AFFILIATION | Based-In | basedin |
| | Citizen-or-Resdent | resident |
| | Other | gpeaffothr |
| PER/ORG AFFIL'N | Ethnic | ethnic |
| | Ideology | ideology |
| | Other | perorgothr |
| PERSONAL-SOC'L | Business | business |
| | Family | family |
| | Other | persocothr |
| PHYSICAL | Located | located |
| | Near | near |

Table 2: Overview of ACE relations with abbreviations used here.

entire collection, and $\mathcal{H}_1$ is a combined criterion function that consists of the ration of $\mathcal{I}_1$ over $\mathcal{E}_1$.

The $\mathcal{I}_2$, $\mathcal{H}_1$ and $\mathcal{H}_2$ criterion functions outperformed single link, complete link and group average on the development data. We use $\mathcal{I}_2$, which performed as well as $\mathcal{H}_1$ and $\mathcal{H}_2$ and is superior in terms of computational complexity (Zhao and Karypis, 2004).

## 5 Experiment

### 5.1 Method

This section describes experimental setup, which uses relation extraction data from ACE 2005 to answer four questions concerning the effectiveness of similarity models based on term co-occurrence and dimensionality reduction for the relation discovery task:

1. Do term co-occurrence models provide a better representation of relation semantics than standard term-by-document vector space?
2. Do textual dimensionality reduction techniques provide any further improvements?
3. How do probabilistic topic models perform with respect to SVD on the relation discovery task?
4. Does one similarity measure (for probability distributions) outperform the others on the relation discovery task?

System configurations are compared across six different data subsets (entity type pairs, i.e., *organisation-geopolitical entity*, *organisation-organisation*, *person-facility*, *person-geopolitical entity*, *person-organisation*, *person-person*) and evaluated following suggestions by Demšar (2006) for statistical comparison of classifiers over multiple data sets.

The dependent variable is the clustering performance as measured by the F-score. F-score accounts for both the amount of predictions made that are true (*P*recision) and the amount of true classes that are predicted (*R*ecall). We use the CLUTO implementation of this measure for evaluating hierarchical clustering. Based on (Larsen and Aone, 1999), this is a balanced F-score ($F = \frac{2RP}{R+P}$) that computes the maximum per-class score over all possible alignments of gold standard classes with nodes in the hierarchical tree. The average F-score for the entire hierarchical tree is a micro-average over the class-specific scores weighted according to the relative size of the class.

### 5.2 Results

Table 3 contains F-score performance on the test set (ACE 2005). The columns contain results from the different system configurations. The column labels in the top row indicate the different representations of relation similarity. The column labels in the second row indicate the dimensional-

| Sem Space | TxD | TxT | TxT | TxT | TxT | TxT |
|---|---|---|---|---|---|---|
| Dim Red'n | None | None | SVD | LDA | LDA | LDA |
| Similarity | Cos | Cos | Cos | KL | Sym | JS |
| ORG-GPE | 0.644 | 0.673 | 0.645 | 0.680 | 0.670 | 0.673 |
| ORG-ORG | 0.879 | 0.922 | 0.879 | 0.904 | 0.900 | 0.904 |
| PER-FAC | 0.811 | 0.827 | 0.831 | 0.832 | 0.826 | 0.820 |
| PER-GPE | 0.595 | 0.637 | 0.627 | 0.664 | 0.642 | 0.670 |
| PER-ORG | 0.520 | 0.551 | 0.532 | 0.569 | 0.552 | 0.569 |
| PER-PER | 0.534 | 0.572 | 0.593 | 0.633 | 0.553 | 0.618 |
| Micro Ave | 0.627 | 0.661 | 0.652 | 0.683 | 0.658 | 0.681 |
| Macro Ave | 0.664 | 0.697 | 0.684 | 0.714 | 0.689 | 0.709 |
| RankAve | 5.917 | 3.083 | 4.250 | 1.500 | 4.000 | 2.250 |

Table 3: F-score performance on the test data (ACE 2005) using agglomerative clustering with the $\mathcal{I}_2$ criterion function.

ity reduction technique used. The column labels in the third row indicated the similarity measure used, i.e. cosine (Cos) and KL (KL), symmetrised KL (Sym) and JS (JS) divergence. The rows contain results for the different data subsets. While we do not use them for analysis of statistical significance, we include micro and macro averages over the data subsets.[11] We also include the average ranks, which show that the LDA system using KL divergence performed best.

Initial inspection of the table shows that all systems that use the term co-occurrence semantic space outperform the baseline system that uses the term-by-document semantic space. To test for statistical significance, we use non-parametric tests proposed by Demšar (2006) for comparing classifiers across multiple data sets. The use of non-parametric tests is safer here as they do not assume normality and outliers have less effect. The first test we perform is a Friedman test (Friedman, 1940), a multiple comparisons technique which is the non-parametric equivalent of the repeated-measures ANOVA. The null hypothesis is that all models perform the same and observed differences are random. With a Friedman statistic ($\chi^2_F$ ) of $21.238$, we reject the null hypothesis at $p < 0.01$.

The first question we wanted to address is whether term co-occurrence models outperform the term-by-document representation of relation semantics. To address this question, we continue with post-hoc analysis. The objective here is to compare several conditions to a control (i.e., compare the term co-occurrence systems to the term-by-document baseline) so we use a Bonferroni-Dunn test. At a significance level of $p < 0.05$, the critical difference for the Bonferroni-Dunn test for comparing 6 systems across 6 data sets is $2.782$. We conclude that the unreduced term co-occurrence system and the LDA systems with KL and JS divergence all perform significantly better than baseline, while the SVD system and the LDA system with symmetrised KL divergence do not.

The second question asks whether SVD and LDA dimensionality reduction techniques provide any further improvement. We observe that the systems using KL and JS divergence both outperform the unreduced term co-occurrence system, though the difference is not significant.

The third question asks how the probabilistic topic models perform with respect to the SVD models. Here, Holm-correct Wilcoxon signed-ranks tests show that the KL divergence system performs significantly better than SVD while the symmetrised KL divergence and JS divergence systems do not.

The final question is whether one of the divergence measures (KL, symmetrised KL or JS) outperforms the others. With a statistic of $\chi^2_F = 9.336$, we reject the null hypothesis that all systems are the same at $p < 0.01$. Post-hoc analysis with Holm-corrected Wilcoxon signed-ranks tests show that the KL divergence system and the JS divergence system both perform significantly better than the symmetrised KL system at $p < 0.05$, while there is no significant difference between the KL and JS systems.

[11] Averages over data sets are unreliable where it is not clear whether the domains are commensurable (Webb, 2000). We present averages in our results but avoid drawing conclusions based on them.

## 6 Discussion

An interesting aspect of using the ACE corpus is the wealth of linguistic knowledge encoded. With respect to named entities, this includes class information describing the kind of reference the entity makes to something in the world (i.e., *specific referential*, *generic referential*, *under-specified referential*) and it includes mention type information (i.e., *names*, *quantified nominal constructions*, *pronouns*). It also includes information describing the lexical condition of a relation (i.e., *possessive*, *preposition*, *pre-modifier*, *formulaic*, , *verbal*). Based on a mapping between gold standard and predicted clusters, we assigned each case a value of 1 or 0 to indicate whether it is a correct or incorrect classification. We then carried out detailed statistical analysis[12] to test for effects of the entity and relation information described above on each system in each domain.

Overall, the effects were fairly small and do not generalise across domains or systems very well. However, there were some observable tendencies. With respect to entity class, relations with *specific referential* entities tend to correlate positively with correct classifications while *under-specified referential* entities tend to correlate negatively with correct classifications. With respect to entity mention type, relations entities that consist of *names* tend to correlate positively with correct classifications while *pronouns* tend to correlate negatively with correct classifications. Though, this is only reliably observed in the PER-GPE domain. Finally, with respect to lexical condition, we observe that *possessive* conditioned relations tend to correlate negatively, especially in the PER-GPE and PER-ORG domains with the PER-PER domain also showing some effect. *Pre-modifier* conditioned relations also tend to correlate negatively in the PER-GPE domain. The effect with *verbally* conditioned relations is mixed. This is probably due to the fact that verbal relations tend to have more words occurring between the entity pair, which provides more context but can also be misleading when the key terms describing the relation do not occur between the entity pair (e.g., the first sentence in Figure 1).

It is also informative to look at overall properties of the entity pair domains and compare this to the system performance. Table 6 contains, for each domain, the F-score of the LDA+KL system, the type-to-token ratio, and the entropy of the relation type distribution for each domain. Type-to-token ratio (TTR) is the number of words divided by the number of word instances and indicates how much repetition there is in word use. Since TTR can vary depending on the size of the text, we compute it on a random sample of 75 tokens from each domain. Entropy can be interpreted as a measure of the uniformity of a distribution. Low entropy indicates a more spiked distribution while high entropy indicates a more uniform distribution. Though there is not enough data to make a reliable conclusion, it seems that the system does poorly on domains that have both a high type-to-token ratio and a high entropy (uniform relation type distribution), while it performs very well on domains that have low TTR or low entropy.

| Domain | Score | TTR | Entrpy |
|---|---|---|---|
| ORG-GPE | 0.680 | 0.893 | 1.554 |
| ORG-ORG | 0.904 | 0.720 | 1.642 |
| PER-FAC | 0.832 | 0.933 | 0.636 |
| PER-GPE | 0.664 | 0.933 | 1.671 |
| PER-ORG | 0.569 | 0.973 | 2.001 |
| PER-PER | 0.633 | 0.867 | 2.179 |

Table 4: System score, type-to-token ratio (TTR) and relation type entropy (Entrpy) for entity pair domains.

## 7 Conclusions and Future Work

This paper presented work on the relation discovery task. We tested several systems for the clustering subtask that use different models of the conceptual/semantic similarity of relations. These models included a baseline system based on a term-by-document representation of term context, which is equivalent to the representation used in previous work by Hasegawa et al. (Hasegawa et al., 2004) and Chen et al. (Chen et al., 2005). We hypothesised that this representation suffers from a sparsity problem and showed that models that use a term co-occurrence representation perform significantly better.

Furthermore, we investigated the use of singular value decomposition and latent Dirichlet allocation for dimensionality reduction. It has been suggested that representations using these techniques are able to model a similarity that is less reliant on

[12] For this analysis, we used the Phi coefficient, which is a measure of relatedness for binomial variables that is interpreted like correlation.

specific word forms and therefore more semantic in nature. Our experiments showed an improvement over a term co-occurrence baseline when using LDA with KL and JS divergence, though it was not significant. We also found that LDA with KL divergence performs significantly better than SVD.

Comparing the different divergence measures for LDA, we found that KL and JS perform significantly better than symmetrised KL divergence. Interestingly, the performance of the asymmetric KL divergence and the symmetric JS divergence is very close, which makes it difficult to conclude whether the relation discovery domain is a symmetric domain or an asymmetric domain like Lee's (1999) task of improving probability estimates for unseen word co-occurrences.

A shortcoming of all the models we will describe here is that they are derived from the basic bag-of-words models and as such do not account for word order or other notions of syntax. Related work on relation discovery by Zhang et al. (2005) addresses this shortcoming by using tree kernels to compute similarity between entity pairs. In future work we will extend our experiment to explore the use of syntactic and semantic features following the frame work of Pado and Lapata (2003). We are also planning to look at non-parametric versions of LDA that address the model order selection problem and perform an extrinsic evaluation of the relation discovery task.

## Acknowledgements

This work was supported by Scottish Enterprise Edinburgh-Stanford Link grant R37588 as part of the EASIE project. I would like to thank Claire Grover, Mirella Lapata, Gabriel Murray and Sebastian Riedell for very useful comments and discussion on this work. I would also like to thank the anonymous reviewers for their comments.

## References

Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. 1994. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595.

Christian Blaschke and Alfonso Valencia. 2002. The frame-based module of the suiseki information extraction system. *IEEE Intelligent Systems*, 17:14–20.

David Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3.

Razvan C. Bunescu and Raymond J. Mooney. 2005. Subsequence kernels for relation extraction. In *Proceedings of the 19th Conference on Neural Information Processing Systems*, Vancouver, BC, Canada.

Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. 2005. Automatic relation extraction with model order selection and discriminative label identification. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1997. Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, Jan.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. Event-based extractive summarization. In *Proceedings of the ACL-2004 Text Summarization Branches Out Workshop*, Barcelona, Spain.

Milton Friedman. 1940. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11:86–92.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.

Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. 2004. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting of Association of Computational Linguistics*.

Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. 2004. Overview of BioCreAtIvE: Critical assessment of information extraction for biology. In *Proceedings of Critical Assessment of Information Extraction Systems in Biology Workshop (BioCreAtIvE)*, Granada, Spain.

Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Buornar Larsen and Chinatsu Aone. 1999. Fast and effective text mining using linear-time document clustering. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA.

Ludovic Lebart and Martin Rajman. 2000. Computing similarity. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *Handbook of Natural Language Processing*, pages 477–505. Marcel Dekker, New York.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD, USA.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Peter V. Marsden and Nan Lin, editors. 1982. *Social Structure and Network Analysis*. Sage, Beverly Hills.

Sebastian Pado and Mirella Lapata. 2003. Constructing semantic space models from parsed corpora. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, D.C., USA.

Barbara Rosario and Marti Hearst. 2004. Classifying semantic relations in bioscience text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):91–124.

Geoffrey I. Webb. 2000. Multiboosting: A technique for combining boosting and wagging. *Machine Learning*, 40(2):159–196.

Min Zhang, Jian Su, Danmei Wang, Guodong Zhou, and Chew Lim Tan. 2005. Discovering relations from a large raw corpus using tree similarity-based clustering. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*.

Ying Zhao and George Karypis. 2004. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55:311–331.