Domain-Neutral Relation Characterisation: Evaluation on Disease-Treatment Data

Ben Hachey BHACHEY@INF.ED.AC.UK
School of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh, EH8 8EY, UK

Abstract

Adapting conventional supervised relation extraction (RE) systems to new domains requires significant effort from annotators and developers. Thus, we propose models for relation characterisation – the subtask of RE that assigns types to extracted relations – that have domain adaptation costs of *zero*. Development experiments on newswire text compare dimensionality reduction techniques and show that a probabilistic model successfully incorporates a larger and more interdependent feature set, outperforming an unreduced model and a linear algebraic model. More importantly, we show that the newswire-optimised model achieves an *f-score* of 0.611 on disease-treatment data. This is nearly identical to performance on the newswire data, proving that the model is capable of generalising to the biomedical domain.

1. Introduction

We evaluate a domain-neutral approach to relation characterisation, a subtask of relation extraction having the goal of labelling relationships between named objects in text (e.g. interactions between proteins, relationships between diseases and treatments). Figure 1 contains two examples from the disease-treatment data used here. In sentence 1 there is a *prevents* relation and in sentence 2 there is a *cured-by* relation.

The genesis and adoption of relation extraction as an active research area owes much, originally, to the MUC Template Element and Template Relation tasks and, latterly, to the ACE Relation Detection/Recognition and BioCreAtIvE II Protein-Protein Interaction tasks (see e.g. Bunescu et al. (2005), Turmo et al. (2006) for history). The systems entered into such competitions are generally based on either rule engineering or supervised machine learning, both of which are expensive and time-consuming to develop. In the case of rule engineering, writing extraction rules requires extensive effort from a language engineer, who must be expert in rule engineering and must at least be competent in the technical language of the target domain. In the case of supervised learning, annotation and tuning of algorithm parameters requires extensive effort from annotators expert in the target domain and from a language engineer. Furthermore, both approaches require annotated development/testing corpora.

1	A two-dose combined *hepatitis A and B* **vaccine** would facilitate immunization programs.
2	These results suggest that con A-induced *hepatitis* was ameliorated by pretreatment with **TJ-135**.

Figure 1. Example sentences with disease entities in italic and treatment entities in bold (Rosario & Hearst, 2004).

The expense of conventional supervised approaches for natural language processing has motivated recent work on partially supervised methods for bootstrapping and domain adaptation.[1] One popular family of approaches for relation extraction starts out with a small seed set for a specific relation and bootstraps a wide-coverage system using unlabelled data, e.g. Brin (1998), Tomita et al. (2006). Other approaches include active learning which presents the annotator with the examples that are deemed most useful for the learning algorithm, e.g. Zelenko et al. (2005).

Partially supervised approaches, however, still require that the developer decide in advance what relation types are important for each new domain. Therefore, Hasegawa et al. (2004) propose a domain-neutral approach, termed relation discovery (RD). RD assumes that named entity recognition (i.e. automatic identification of named objects such as *people*, *organisations*, *diseases*, *treatments*) can be performed. Then, domain-neutral methods are applied to the subtasks of (1) identifying entity pairs in the text that form a relation and (2) characterising relations by annotating them with a type (e.g. *prevents* and *cured-by* for the examples in Table 1).[2]

[1] See e.g. the CoNLL-2007 shared task on dependency parsing, which includes a domain adaptation task: `http://depparse.uvt.nl/depparse-wiki/SharedTaskWebsite`.

[2] RD can be seen as an extension of data mining approaches which address only relation identification, e.g. Conrad and Utt (1994), Wren (2004).

1	Build vector space
1.1	*Extract features for relations*
1.2	*Build relation-by-term vector space*
2	Build similarity models
2.1	*Do dimensionality reduction*
2.2	*Build similarity matrix*
3	Perform clustering

Figure 2. Generalised framework for rel. characterisation.

The focus of this paper is on RD subtask 2 (domain-neutral relation characterisation), where the goal is to induce a partition over the data that groups relations by type through the application of unsupervised clustering techniques.[3] Related work here has focused on clustering models (Hasegawa et al., 2004; Zhang et al., 2005; Hachey, 2006; Chen et al., 2006) and the genericity claim has not been explicitly addressed by performing an evaluation across different domains. In Section 2, we introduce three models with respect to a generalised framework. Next, in Section 3, we report experiments comparing models and feature combinations. Finally, in Section 4, we present the genericity experiment which applies a newswire-optimised model directly to biomedical disease-treatment data.

2. Generic Relation Characterisation

Work to date on the RD relation characterisation subtask has focused on the primary modelling problem of feature representations for clustering. As such, some fairly complex models have been devised which further motivate an explicit evaluation of genericity. Figure 2 contains a generalised framework for domain-neutral relation characterisation that will be discussed in more detail in the rest of this section with respect to the framework of the experiments presented here.

The first step is to build the vector space. This consists of building feature vectors for relation instances (entity pairs) based on the textual context. In the current work, we use feature sets that are combinations of intervening word features (W), entity word features (E), and features extracted from the shortest dependency path between the entity pair (D). Using features derived from dependency paths for RD is a novel divergence from Zhang et al. (2005) and Chen et al. (2006) who use features derived from phrase structure parses for RD. The exact configuration of the feature sets was optimised on the development data resulting in the following approach. For entities that are nominal (e.g. 'surgery', not 'Amifostane'), we collect words, stopping all function words except *where*. For dependency path features, we parse with Minipar and identify the most direct path between the entity pair (Lin & Pantel, 2001; Snow et al., 2006). E.g., for the sentence "Amifostine has also been shown to *stimulate haematopoietic stem cells* and has been *investigated* as a *therapy* for *patients* with myelodysplastic syndrome," the path between treatment 'Amifostine' and disease 'myelodysplastic syndrome' is:

While the W features include all non-stop tokens between the entities (italicised in sentence), the D features skip the first conjunct, taking in only the tokens *investigated*, *therapy* and *patients* and the dependency relations *obj*, *as*, *for* and *with*.

The second step in the generalised framework is to create the similarity matrix that will be used for clustering. Our primary modelling contribution is a comparative evaluation of different dimensionality reduction techniques for language data. Dimensionality reduction for linguistic vector spaces is motivated by a number of studies showing that it represents a type of semantic similarity that is more linguistic in nature (Landauer et al., 1998). Singular value decomposition (SVD) is a technique from linear algebra (Berry et al., 1994), which has been applied to a number of NLP and cognitive modelling tasks. Latent Dirichlet allocation (LDA) is a probabilistic technique analogous to SVD (Blei et al., 2003), whose contribution has not been as thoroughly explored. In the current work, we compare SVD and LDA under optimal conditions, where the development data has been used to tune model parameters including dimensionality and LDA hyperparameters (see Hachey (2006), Section 4.2). By contrast, Chen et al. (2006) demonstrate a non-parametric approach based on spectral clustering. To complete the second step in the framework, we build a similarity matrix. For the unreduced and SVD-reduced models, we do this by computing the cosine similarity between the entity pair vectors. For LDA, as the output is a probability distribution over latent topics, we use KL divergence to compare entity pair vectors. Therefore, we also tune a constant C for divergence-to-similarity conversion where $sim(x) = C - KL(x)$.

Finally, the third step in the generalised framework for domain-neutral relation characterisation is to induce a partition over the data that groups entity pairs by relation type. We follow the convention in the RD literature by using hierarchical agglomerative clustering, a choice which is motivated by the fact that it is not known in advance how many clusters there should be in a new domain (Hasegawa et al., 2004). We also follow previous work in using the $\mathcal{I}_2$ criterion function as a measure of the optimality of the clustering solution. The $\mathcal{I}_2$ criterion function maximises the similarity between each instance and the centroid of the cluster it is assigned to and was found by Hachey (2006) to outperform other criterion functions for RD.

[3] Gold standard named entity recognition and relation identification are used here to isolate the performance of relation characterisation.

3. Development Experiments

3.1. Materials

Following Chen et al. (2006) and Hachey (2006), we use annotated relation extraction data for our development experiments. This allows unbiased evaluation compared to an established gold standard. We use a subset of the automated content extraction (ACE)[4] 2004 newswire and broadcast news data, which consists of subdomains based on four entity types: persons (PER), organisations (ORG), geographical/social/political entities (GPE), and facilities (FAC). Six subdomains are used: ORG-GPE, ORG-ORG, PER-FAC, PER-GPE, PER-ORG, and PER-PER. While space restrictions prevent full description, the ORG-GPE subdomain has the following relations (counts): *based-in* (54), *subsidiary* (27), *located* (15), *other* (3) and PER-FAC has: *located* (127), *owner* (14), *near* (4). See Hachey (2006), Tables 1 and 2 for details.

3.2. Method

This section describes the experimental setup, which aims to answer the following questions: *Which model – unreduced, SVD-reduced or LDA-reduced – is best? Which combination of features is best?* Model configurations are compared across the six different subsets of ACE 2004 described in the previous section (3.1).

Following Chen et al. (2006), we report an *f-score* metric based on a one-to-one mapping Ω between the partition over the relations given by our clustering system and the gold standard partition from the annotated data. Then, precision and recall for a cluster i are defined as:

$$p(i) = \frac{m_{i,\Omega(i)}}{m_i} \qquad r(i) = \frac{m_{i,\Omega(i)}}{m_{\Omega(i)}}$$

where $\Omega(i)$ is the gold standard class from the one-to-one mapping, $m_{i,\Omega(i)}$ is the size of the intersection between i and $\Omega(i)$, m_i is the size of cluster i, and $m_{\Omega(i)}$ is the size of class $\Omega(i)$. The overall score is calculated as the micro-average of individual cluster scores. The *f-score* is calculated as $2pr/(p+r)$.

3.3. Results

Results are given in Table 1 where rows correspond to different feature sets and the first three columns correspond to clustering with an unreduced vector space (Cl:None), with a SVD-reduced space (Cl:SVD), and with a LDA-reduced space (Cl:LDA). The fourth column (UB:ME) corresponds to an upper bound obtained by performing 10-fold cross-validation with a standard supervised learning algorithm (maximum entropy[5]) and using the resulting classification to define a partition over instances that is evaluated in the same way as the clustering output.

	Cl:None	Cl:SVD	Cl:LDA	UB:ME
W	0.465	0.529	0.598	0.689
D	0.484	0.542	0.561	0.659
WE	**0.538**	0.546	0.580	0.700
ED	0.462	0.521	0.572	0.745
WED	0.458	**0.547**	**0.604**	0.750

Table 1. F-score results for different models across feature sets. Lower bound (random partition) is 0.351.

First, we observe that the best unreduced model is obtained with intervening word and entity word feature sets while the best SVD- and LDA-reduced models are obtained when all features are used. This suggests that dimensionality reduction acts to smooth the relation vectors in a way that allows them to incorporate a larger and more interdependent feature set. Comparing these models to the lower bound (random partition) using a paired Wilcoxon signed-ranks test, we find that they are all significantly better ($p < 0.05$). More importantly, though, we observe that the best model is Cl:LDA, which is significantly better than the unreduced model ($p = 0.0156$) and nearly significantly better than the SVD-reduced model ($p = 0.0781$).

4. Genericity Experiment

4.1. Materials

For genericity testing we adapt the BioText disease-treatment (DT) data (Rosario & Hearst, 2004). This data consists of 3655 sentences from Medline 2001 with entities annotated at token level and relations annotated at the sentence level. We use only sentences where relations can be automatically mapped to entities (i.e. those with one disease and one treatment entity). From these sentences, we choose instances for clustering using the same criteria applied to the ACE data (see Hachey (2006), Section 4.1). Half of this data is used for the current evaluation (saving the remainder for further development and testing). The relation types in the resulting data set are given in the middle column of Figure 3 with a brief description. The number of relation instances per relation type is given in the left column.

Thus, the final DT data set used here has 4 relation types which is comparable to the mean number across entity pair domains in the ACE development data – μ: 5.17, *95% CI*: (3.46, 6.88). The total number of relation instances (248) is also similar to the mean in ACE – μ: 190, *95% CI*: (90.47, 289.53). The distribution across relation types is more skewed than the mean ACE domain, though it is not a complete departure as the PER-FAC (see Section 3.1) domain in ACE is similarly skewed.

[4] ACE is the NIST shared task on information extraction. See `http://www.nist.gov/speech/tests/ace/`.

[5] `http://www.cs.utah.edu/~hal/megam/MEGA`

218	Cures	*Treatment* cures *disease*
15	Prevents	*Treatment* prevents *disease*
10	Vague	Unclear relationship
5	Side Effect	*Disease* result of *treatment*

Figure 3. Disease-treatment relation counts and types.

LB:Rand	Cl:LDA	UB:ME
0.315	0.611	0.869

Table 2. Disease-treatment evaluation f-score results.

4.2. Method

This section describes the experimental setup which addresses the question: *Does generic relation characterisation generalise to the biomedical domain?* We use the optimised model from the ACE development data as described in the previous Section (3.3), i.e. the Cl:LDA model with intervening words, entity words, and dependency path feature sets. This is applied to the DT data without modification and evaluated using the *f-score* based on a one-to-one mapping between system clusters and gold standard classes (Section 3.2).

Note that the evaluation presented here is somewhat contrived in that development data should be as representative as possible of the data to which the system will eventually be applied. Therefore, when building a domain-neutral system, one would tune across as many different domains as possible. However, for assessing genericity, the setup here is illustrative in that it allows us to test the somewhat extreme condition of tuning and evaluating in completely different domains.

4.3. Results

Table 2 contains *f-score* performance on the DT data. The columns contain results for the lower bound (LB:Rand), the newswire-optimised model (Cl:LDA), and the upper bound (UB:ME). The lower bound is a random partition and the upper bound is derived from 10-fold cross-validation with maximum entropy. The Cl:LDA model performs 0.296 points better than the random lower bound, which corresponds to a 53.4% reduction in error rate with respect to the upper bound. To answer our experimental question, we consider this with respect to performance on the development data.

First, we note that the *f-score* is very close to the mean across ACE domains (0.604). Secondly, we compare the reduction in error rates. Again, the model does well on the DT data at 53.4%, which is not quite as good as the Cl:LDA model on the development data (63.4%) but better than the unreduced and SVD-reduced models (46.9% and 49.1% respectively). Thus, we conclude that our approach to generic relation characterisation is capable of generalising to the biomedical domain with an adaptation cost of zero.

5. Discussion and Future Work

We described results for domain-neutral relation characterisation which explicitly demonstrate for the first time that the approach generalises across domains. Applications include summarisation, where relations can be used as a conceptual representation for sentence extraction or used directly to generate biographical sketches for named entities. Relation discovery could also be integrated into a fully bottom-up bootstrapping approach to annotation where user feedback with respect to RD output would feed into partially supervised approaches. In future work, we plan to further test the genericity of our models by evaluating on protein-protein interaction data and on data from a third domain. We also plan to test the genericity of domain-neutral relation identification.

References

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1994). Using linear algebra for intelligent information retrieval. *SIAM Review*, *37*, 573–595.

Blei, D., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of ML Research*, *3*, 993–1022.

Brin, S. (1998). Extracting patterns and relations from the world wide web. *1st WebDB*. Valencia, Spain.

Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., & Wong, Y. W. (2005). Comparative experiments on learning information extractors for proteins and their interactions. *AI in Med.*, *33*, 139–155.

Chen, J., Ji, D., Tan, C. L., & Niu, Z. (2006). Unsupervised relation disambiguation with order identification capabilities. *COLING/ACL*. Sydney, Australia.

Conrad, J. G., & Utt, M. H. (1994). A system for discovering relationships by feature extraction from text databases. *17th SIGIR*. Dublin, Ireland.

Hachey, B. (2006). Comparison of similarity models for the relation discovery task. *COLING/ACL WS Linguistic Distances*. Sydney, Australia.

Hasegawa, T., Sekine, S., & Grishman, R. (2004). Discovering relations among named entities from large corpora. *42nd ACL*. Barcelona, Spain.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Lin, D., & Pantel, P. (2001). Discovery of inference rules for question answering. *Natural Lang. Eng.*, *7*, 343–360.

Rosario, B., & Hearst, M. (2004). Classifying semantic relations in bioscience text. *42nd ACL*. Barcelona, Spain.

Snow, R., Jurafsky, D., & Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. *COLING/ACL*. Sydney, Australia.

Tomita, J., Soderland, S., & Etzioni, O. (2006). Expanding the recall of relation extractin by bootstrapping. *EACL WS Adaptive Text Extraction and Mining*. Trento, Italy.

Turmo, J., Ageno, A., & Català, N. (2006). Adaptive information extraction. *ACM Computing Surveys*, *38*.

Wren, J. D. (2004). Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics*, *5*.

Zelenko, D., Aone, C., & Tibbetts, J. (2005). Trainable evidence extraction system. *IA'05*. McLean, VA, USA.

Zhang, M., Su, J., Wang, D., Zhou, G., & Tan, C. L. (2005). Discovering relations from a large raw corpus using tree similarity-based clustering. *2nd IJCNLP*. Jeju Island, Korea.