MDL-based Acquisition of Substitutability Relationships between Discourse Connectives

Ben Hutchinson
University of Edinburgh
School of Informatics
2 Buccleuch Place, Edinburgh, EH8 9LW, UK
B.Hutchinson@sms.ed.ac.uk

Abstract

Knowledge of which lexical items convey the same meaning in a given context is important for many Natural Language Processing tasks. This paper concerns the substitutability of discourse connectives in particular. This paper proposes a data-driven method based on a Minimum Description Length (MDL) criterion for automatically learning substitutability of connectives. The method is shown to outperform two baseline classifiers.

1 Introduction

Discourse connectives are words or phrases which explicitly signal coherence relations in texts, e.g. *however*, *even though* and *whereas*. Studies of connectives have provided insights into natural language discourse, however formal semantic analysis of connectives is a difficult task, and most connectives have not been the subject of much study. The empirical study of the substitutability of discourse connectives can inform and complement theoretical analyses [Knott, 1996]. The substitutability of discourse connectives is also important for Natural Language Processing applications such as Natural Language Generation and text simplification [Moser and Moore, 1995; Siddharthan, 2003].

This paper investigates acquiring substitutability relationships between connectives automatically. Previous attempts to do so have been hindered by the relatively low prior likelihood of two connectives being substitutable [Hutchinson, 2005]. We introduce an MDL-based approach that addresses this.

2 Substitutability of discourse connectives

Example (1) contains the connective *seeing as*. However, *because* can also be used to achieve the same discourse aims, i.e. it is **substitutable** [Knott, 1996].

(1) **Seeing as/because** we've got nothing but circumstantial evidence, it's going to be difficult to get a conviction.

However, *because* it is not always substitutable for *seeing as*, and the converse does not hold either. In general, there are five possibilities for a connective X's relationship to another connective Y:

- X is a SYNONYM of Y if X can always be substituted for Y, and vice versa.
- X and Y are EXCLUSIVE if neither can ever be substituted for the other.
- X is a HYPONYM of Y if Y can always be substituted for X, but not vice versa.
- X is a HYPERNYM of Y if X can always be substituted for Y, but not vice versa.
- X and Y are CONTINGENTLY SUBSTITUTABLE if each can sometimes, but not always, be substituted for the other.

We would like to automatically predict which of these relationships holds, for any given pair of connectives.

3 An MDL-based model of substitutability

We will refer to a consistent set of relationships between connectives as a *taxonomy*. Our model of taxonomies exploits the following two observations. Firstly, it is unlikely that *every* connective should be EXCLUSIVE with every other one. Secondly, there are logical constraints on substitutability. For example, if A is a HYPONYM of B and B is EXCLUSIVE with C, then A must also be EXCLUSIVE with C.

Our modelling of substitutability is within the Minimum Description Length (MDL) framework. In our case, we wish to evaluate models representing substitutability relationships between connectives. For a taxonomy $\mathbb{T}$ and data $data$, the total description length is given by:

$$L(\mathbb{T}, data) = L(\mathbb{T}) + L(data|\mathbb{T}) \quad (2)$$

We will exploit the fact that description lengths L can be related to probabilities P via the equation $L = -\log_2 P$.

Calculating the prior: We will use the notation $\langle rel, X, Y\rangle \in \mathbb{T}$ to indicate that in the taxonomy $\mathbb{T}$ connectives X and Y are in relationship rel. We calculate $P(\mathbb{T})$ using the following multinomial model:

$$P(\mathbb{T}) = M \prod_{\langle rel, X, Y\rangle \in \mathbb{T}} P(rel) \quad (3)$$

where (i) $P(rel)$ is the prior probability of two connectives being in the relationship rel, which will be estimated empirically, and (ii) M is a multinomial coeffient which ensures that the most likely taxonomy contains numbers of each pairwise substitutability relationship in proportion to their prior probabilities.

The multinomial model (3) is defined over all sets of pairwise substitutability relationships, but we are only interested in calculating the probabilities of consistent sets. However this does not affect the relative likelihoods of taxonomies with the same number of connectives, so for the purposes of the experiments below this effect can be ignored.

Estimating the posterior probability: Previous work has found a correlation between substitutability and distributional similarity [Hutchinson, 2005]. Therefore, the data that our model of taxonomies aims to explain are the Kullback-Leibler (KL) divergences, $D(p||q)$, between pairs of connectives. That is, we assume that the data to be explained is:

$$data \equiv \{D(X||Y) : \langle rel, X, Y\rangle \in \mathbb{T}\}$$

To estimate the probability of the data, we assume that the likelihood of observing a given distributional divergence $D(X||Y)$ between X and Y is dependent only on the substitutability of X and Y.

$$P(data|\mathbb{T}) \approx \prod_{\langle rel,X,Y\rangle\in\mathbb{T}} P(D(X||Y)|\langle rel, X, Y\rangle) \quad (4)$$

To estimate each of the multiplicands in (4), we use Gaussian models of the distributional divergences corresponding to each substitutability relationship, i.e.

$$P(D(X||Y)|\langle rel, X, Y\rangle) \propto n(D(X||Y); \mu_{rel}, \sigma_{rel}) \quad (5)$$

where μ_{rel} and σ_{rel} and the mean and standard deviation of KL divergences of all pairs of connectives in relationship rel.

4 Experiment

Task: Our task is to predict the substitutability relationships in a manually constructed taxonomy $\mathbb{T}$ of connectives. We iteratively remove a single connective from $\mathbb{T}$ and attempt to re-insert it in its original position. To do this, we insert it in the position that minimises the description length.

Methodology: An existing taxonomy containing 80 discourse connectives was used as a gold standard [Hutchinson, 2005]. Fixed width beam search (width=1000) is used to search for the taxonomy with minimum description length. Co-occurrences with other discourse connectives were used as the distributional representations.

Parameter tuning: Due to unrealistic independence assumptions in Equation (4), our model underestimates the value of $P(data|\mathbb{T})$. As a result, it overestimates $L(data|\mathbb{T})$. To counterbalance this effect, we weight $L(data|\mathbb{T})$ by a parameter $\lambda \in (0, 1]$. Using half the pairs of connectives as a validation set, the setting $\lambda = 0.1$ was found to give good results.

Evaluation metrics: Two evaluation metrics were used: overall accuracy in predicting substitutability relationships, and the amount of information (measured in bits) of the correctly classified instances [Kononenko and Bratko, 1991].

Baselines: Two baseline classifiers were constructed for comparison with the MDL-based model. The first assigned all pairs to the most frequent relationship, namely EXCLUSIVE. The second baseline classifier assumed that the new connective would be a SYNONYM of one other connective X already in the taxonomy. All other relationships involving the new connective were thus identical to X's. When evaluated on all of connectives, these classifiers had accuracies of 69.9% and 70.7%, and correctly learnt 2280 and 3333 bits of information, respectively.

Results: On the 50% of pairs of connectives not in the validation set, the MDL-based classifier had an accuracy of 73.8%, which is significantly greater than both baselines. The improvement is even greater on the information theoretic metric: correct classifications had 1928 bits of information, from which we can extrapolate a performance of about 3856 bits on the set of all pairs of connectives.

5 Discussion and conclusions

We have introduced a statistical model of the lexicon based on the Minimum Description Length principle. A multinomial prior assigns the greatest probability when the frequencies of pairwise relationships are in proportion to the prior likelihoods of those relationships. The model is applied to extending a taxonomy representing substitutability relationships between discourse connectives, and gives better performance than two classifiers that do not take into account global properties of the lexicon. The model aims to explain the distributional similarity of pairs of connectives, as measured by the Kullback-Leibler divergence function. In future, we will explore the use other distributional similarity functions, as these provide other perspectives of the data that may prove useful for learning substitutability. The utility of including other types of co-occurrence data in our distributional representations also remains to be explored.

References

[Hutchinson, 2005] Ben Hutchinson. Modelling the substitutability of discourse connectives. In *Proceedings of the 43rd Annual Meeting of The Association for Computational Linguistics*, Ann Arbor, USA, 2005.

[Knott, 1996] Alistair Knott. *A data-driven methodology for motivating a set of coherence relations*. PhD thesis, University of Edinburgh, 1996.

[Kononenko and Bratko, 1991] Igor Kononenko and Ivan Bratko. Information-based evaluation criterion for classifier's performance. *Machine Learning*, 6:67–80, 1991.

[Moser and Moore, 1995] M. Moser and J. Moore. Using discourse analysis and automatic text generation to study discourse cue usage. In *Proceedings of the AAAI 1995 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 92–98, 1995.

[Siddharthan, 2003] Advaith Siddharthan. Preserving discourse structure when simplifying text. In *Proceedings of the 2003 European Natural Language Generation Workshop*, pages 103–110, 2003.