Supervised Recognition of Entailment Between Patterns

Julien Kloetzer　Stijn De Saeger　Kentaro Torisawa
Motoki Sano　Jun Goto　Chikara Hashimoto　Jong Hoon Oh
NICT - Information Analysis Laboratory
{julien,stijn,torisawa,msano,goto-j,ch,rovellia}@nict.go.jp

1 Introduction

In this paper we present a supervised recognition method for entailment between binary lexico-syntactic patterns such as *X is the capital of Y* and *X is in Y*. Recognizing entailment relations between patterns is useful for applications such as question answering, which is our main motivation in this work.

Since sentences entailing each other are natural paraphrases, entailment is closely related to paraphrasing. Many researchers have successfully used unsupervised distributional similarity based methods for paraphrase acquisition [4, 6, 1], and our own experience with NICT's spoken question answering system Ikkyu [7] [1] confirms their effectiveness.

If Ikkyu could also detect that *X is the capital of Y* entails *X is in Y*, it would be able to answer the question "*Where is Paris?*" from the information that "*Paris is the capital of France*". However, *X is the capital of Y* and *X is in Y* are not strict paraphrases, and indeed their distributional profiles exhibit large differences. Ikkyu's current paraphrasing engine is based on distributional similarity between patterns, and so is highly sensitive to such differences. This is the reason Ikkyu currently cannot exploit the information that "*Paris is the capital of France*" to answer the question "*Where is Paris?*". By adding an accurate and robust entailment recognition module that can recognize entailment pairs even with large differences in distributional profile, we aim to further improve Ikkyu's recall.

In this work we explore a supervised method for entailment recognition that uses both distributional similarities and surface/syntactic features. We show that this supervised approach yields better performance than state-of-the-art unsupervised methods, like DIRT [4] or the scoring method from [2], and than supervised methods that only consider surface similarity like [5] for all types of pattern pairs, even those with very low surface similarity (i.e. sharing no content words).

Our approach is targeted at Japanese but is easily applicable to other languages. We present in Section 2 a description of the resources and the features used, and in Section 3 our experimental methodology and a discussion of our results.

[1] http://www2.nict.go.jp/x/x161/index.html

2 System description

In Section 2.1 and 2.2 below we present the resources and features used in this study.

2.1 Data Description

Class Dependent Patterns We extracted lexico-syntactic patterns from a corpus of 600 million web pages parsed with KNP (Kurohashi-Nagao Parser) [2]. In this work, patterns consist of words on the path of dependency relations connecting two nouns in a sentence. We obtained 70 million unique patterns and their co-occurring noun pairs from our corpus.

Following [1], we consider the entailment relation between patterns to be dependent on the nouns filling the patterns' argument slots: if X is a place name and Y is a drink then "*X's Y is delicious*" entails "*(one) can drink Y at X*". Hence we treat entailment as a relation between *class dependent* patterns, i.e. patterns whose noun arguments are restricted to certain semantic classes. These semantic classes are obtained using the EM-based noun clustering method proposed in [3], which we used to cluster 1 million nouns into 500 semantic word classes.

Given a pattern p and a semantic class pair cp, let $np(p, cp)$ be the set of noun pairs from cp co-ocurring with pattern p. In the rest of this paper and unless stated otherwise, nouns or noun pairs co-occurring with a pattern will always be considered to be from a given class pair, so we will note $np(p, cp)$ as $np(p)$.

Alagin Databases The *Advanced Language Information Forum*[3] (Alagin) provides lexical resources for Japanese. Among them we used: **(1)** databases of verb entailment and non-entailment (Alagin resource ID A-2), **(2)** databases of allographic words (ID A-7), and **(3)** databases of synonyms, antonyms and part-of word pairs (ID A-9).

2.2 Features description

Surface features [5] describes several similarity measures based on surface features used for training an SVM to recognize entailment between sentences. We used their feature set as a basis for our classifier.

[2] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP
[3] http://www.alagin.jp/

Table 1: Features

Feat. type	Data type	feature	count
Surface	token, token stem and token POS tags 1,2,3-grams, nouns, stem nouns, verbs, stem verbs	11 similarity measures	143
Surface	pattern length	ratio, invert ratio, raw values	4
Surface	negative, token 1,2-grams	presence	$\sim$ 56000
Dis. similarity	co-occurring noun pairs, nouns	5 similarity measures	15
Dis. similarity	distribution of co-occurring nouns over POS tags	5 similarity measures, raw values	178
Dis. similarity	class pair	presence	$\sim$ 500
Databases	verb pairs, noun pairs, word pairs	presence	52

Following the approach in [5], we build for each pattern the following sets: **(1)** original token n-grams (for n=1, 2, 3), **(2)** stem forms of **(1)**, **(3)** part-of-speech (POS) of **(1)**, **(4)** nouns contained in the pattern, **(5)** stems of these nouns, **(6)** verbs contained in the pattern, and **(7)** stems of these verbs.

We use the above bag-of-word representations of each pattern to compute pattern similarities. Following [5] we used the following measures: cosine distance, dice coefficient, Jaccard coefficient, Euclidian distance, Jaro distance, Levenshtein distance, Manhattan distance and matching coefficient. To these we added the following measures:

Discounted Jaccard coefficient

$$DisJaccard(U,V) = \frac{|U \cap V|}{|U \cup V|} * \frac{|U \cap V|}{|U \cap V| + 1}$$

Overlap ratios

$O_U(U,V) = \frac{|U \cap V|}{|U|}$ and $O_V(U,V) = \frac{|U \cap V|}{|V|}$

Other surface features include: **(1)** the presence of a negation in each pattern, **(2)** the length of each pattern, **(3)** the length ratio between both patterns and its invert, and **(4)** the presence/absence of each token 1-gram or 2-gram in each pattern.

Distributional similarities Using the data presented in Section 2.1, we used as features for a pattern pair (p, q): **(1)** the Jaccard coefficient and **(2)** the discounted Jaccard coefficient between $np(p)$ and $np(q)$, **(3)** the overlap ratios of these sets, and **(4)** the size of their intersection. We compute the same similarities for the nouns filling the patterns' respective argument slots.

We also computed the distribution of the nouns filling each patterns' respective argument slot over the POS tags given by JUMAN [4] (42 major and minor POS tags). We then computed the above similarity measures (1), (2), (3) and (4) for these distributions to use as features. Finally, we included POS tag frequencies over the nouns and the class pair in question directly as features.

[4] `http://nlp.ist.i.kyoto-u.ac.jp/index.php?cmd=read&page=JUMAN`

Lexical Resources For each Alagin database we signal as a feature the presence of a word pair from this database when for some pattern pair (p, q), p contains the first and q the second word of the pair (and vice-versa). We do the same for stemmed versions of the words in each pattern.

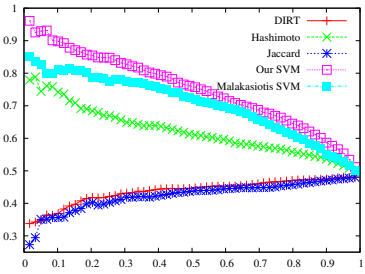
3 Evaluation

We compare the performance of our method to unsupervised baselines based on distributional similarity, and to one other supervised approach based on surface/syntactic similarity. We show (a) that distributional similarity based methods like Jaccard or DIRT cannot cope well with pairs with large differences in corpus frequency, and (b) that our approach obtains a precision of 80% for at least 10% recall, whether patterns share content words or not. We also show (c) that all three feature types contribute to the overall performance of our classifier.

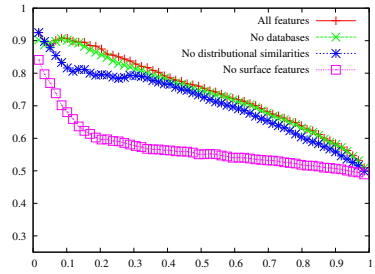
Evaluation Data Given a pair of class dependent patterns, our task is to judge whether the first pattern (*entailing*) entails the second (*entailed*). Since the *entailed* pattern usually has a broader meaning than the *entailing* one, we limited ourselves to pattern pairs where the *entailed* one is a frequent pattern. We define the 25 most frequent patterns of each class pair (co-occurring with at least 10 unique noun pairs) as *reference patterns*, and gathered about 200,000 of them. These represent the most typical relations between nouns in a class pair. For instance in the class pair *(City, Country)* reference patterns include *X is in Y* and *X is close to Y*.

To build our test set we extracted around 10,000 class dependent pattern pairs this way: we first selected a random pattern p, and then a random reference pattern which shares co-occurring noun pairs with p. Our data covers 216 class pairs, with at most 100 pattern pairs per class pair. We then asked 4 annotators to judge the entailment relation between the patterns. Because our semantic classes are obtained using noun clustering they do not have meaningful labels, so for each pattern pair we included three noun pairs co-occurring with the *entailing* pattern in the corpus, as representative for the classes. From early experiments, providing the annotators with more than three noun pairs did not significantly

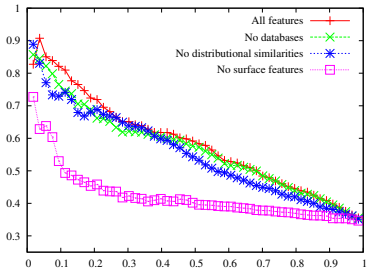
Figure 1: Precision-recall: algorithms comparison and ablation tests



(a) Precision/Recall on all algorithms



(b) Precision/Recall - ablation test



(c) Precision/Recall on **other** pattern pairs - ablation test

alter their judgement. Removing pairs the annotators marked as ungrammatical gave around 9500 pattern pairs. The inter annotator agreement (kappa) was of 60.4, representing a substantial agreement.

Evaluation Setting We used an SVM classifier to perform 10-fold cross validation on the above test set and ranked all pattern pairs according to their classification score. Each fold was built such that its class pairs do not overlap with any other fold, to check our classifier's ability to generalize to new class pairs not seen during training. For training we use TinySVM[5] with a polynomial kernel of degree 2.

We compare our classifier to the methods below.

1. **Jaccard based similarity**: For this baseline we used previously defined similarity measure DisJaccard(np(p),np(q)) (Section 2.2) to score each pattern pair (p, q). This is the similarity measure used by our QA system Ikkyu for paraphrase acquisition. We stress that the score of this baseline on this particular data set is not indicative of Ikkyu's *paraphrase* acquisition performance. The test data in this work consists of pattern pairs where the entailed pattern is explicitly selected from a list of high-frequency patterns. As a result, our test data contains many pattern pairs with a very different distributional profile. Ikkyu's paraphrase acquisition method however selects as paraphrases those patterns that are most distributionally similar to the question's pattern from a list of 70 million candidate patterns. Its top paraphrases are therefore unlikely to include many of the pattern pairs in this study's test data.

2. **DIRT** (Lin and Pantel, [4]) is an unsupervised scoring method to detect inference rules. It is based on distributional similarity: the more significant noun pairs (in terms of mutual information between pattern and noun pair) two patterns share, the more similar they are.

3. **Hashimoto et al.** [2] proposed an unsupervised method for recognizing verb entailment. Their score, adapted here to pairs of patterns, is a directional distributional similarity measure based on a conditional probability.

4. **Malakasiotis et al.** [5] use an SVM with surface features to detect entailment; these include surface similarity measures between the patterns, the presence of a negation, and the length ratio of the patterns, all presented in Section 2.2.

[5] `http://chasen.org/ taku/software/TinySVM/`

3.1 Discussion

For each tested method we ranked the evaluated samples by their score and measured the performance by precision/recall. The precision/recall graph for all methods can be found in Figure 1(a). The results show that our approach outperforms all other methods on both precision and recall. Using this classifier and our corpus data, we should be able to obtain around 500 million pattern pairs with an entailment relation with a precision of 90%. It also confirms that distributional similarity based methods like Jaccard and DIRT work poorly on pattern pairs that differ widely in frequency, like our test data.

We found the pattern pairs in our test data naturally fall into three categories: (a) **similar** pattern pairs, where both patterns share at least one content word, or the *entailed* pattern can be obtained from the *entailing* one by removing some words, (b) **X's Y** pattern pairs where the *entailed* pattern is *X's Y*, and (c) **other** pairs, self-explained category. These categories represent respectively 24.5%, 33.4% and 42.1% of the data. Table 2 shows examples ranked high by our classifier for each of them.

Figure 2 shows the precision/recall graphs obtained by restricting the ranked samples to these three categories (without retraining the classifiers). Our method outperforms every other method on all three categories, though in the case of **similar** pattern pairs the difference with the surface similarity based SVM of the Malakasiotis et al. is minimal, as can be expected. This confirms that our approach can detect entailment relations with a high precision, regardless of the category of the pattern pair. Particularly for pattern pairs that do not share any content words (**other** and **X's Y** pairs) the performance gain

Figure 2: Precision/Recall on all tested algorithms for similar pairs (left), "X's Y" pairs (center) and other pairs (right)

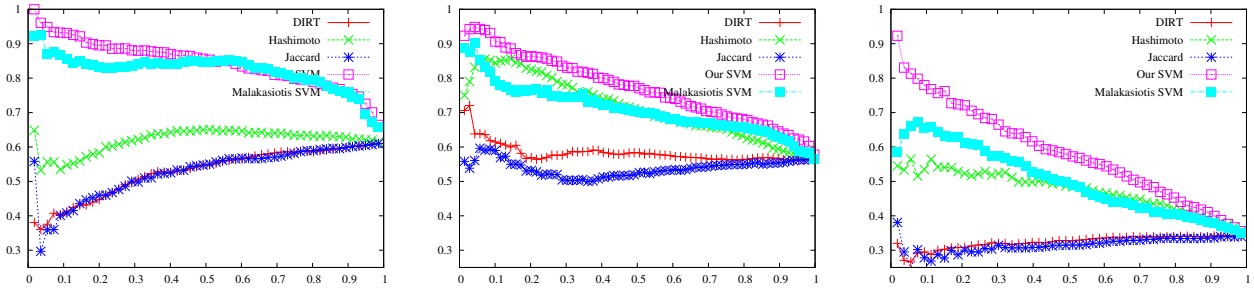


Table 2: Entailment examples

Rank (out of 4415)	Pattern pair	Example noun pair	category
6	X を行った Y - X を行う Y (a) Y who presented X - (a) Y presenting X	研究発表 / 学生会員 research presentation / member student	similar
166	X は Y がいる - X の Y there are Y at (the) X - X's Y	講義 / 出席者 lecture / attendant	X's Y
413	X の販売を行う Y - X を扱っている Y (a) Y selling X - (a) Y dealing with X	ＤＶＤ等 / 会社 dvds etc. / company	other
794	X から Y を受ける - X の Y get (a) X from Y - X's Y	英国人 / 講義 English person / lecture	X's Y
1071	Y が X を入れすぎる - Y が X を入れる Y shows too much X - Y shows X	気合 / コーチ spirit / coach	similar
1850	Y で X をする - Y で X を開く doing a X in Y - opening a X in Y	コンサート / 海外 concert / foreign country	other

over the compared methods is quite large.

We also performed ablation tests by removing each type of feature (surface features, distributional similarities, and Alagin databases based features) in turn, the results of which can be seen in Figure 1(b). Surface features and distributional similarity based features are clearly important in all cases, and databases based features specifically for pattern pairs not sharing content words as shown by restricting the results to **other** pairs (Figure 1(c)).

4 Conclusion

We presented here our approach to detect entailment relations between patterns inside a given class pair using a supervised classifier based on surface features of the patterns, distributional similarities and lexical resources. Our classifier outperforms every compared baseline method, even for pattern pairs with no common content words. We have also shown that all three types of features are necessary to obtain this result. We are considering the possibility of releasing our data (both annotated pattern pairs and pattern pairs classified by our method) through Alagin.

References

[1] S. De Saeger, K. Torisawa, J. Kazama, K. Kuroda, and M. Murata. Large scale relation acquisition using class dependent patterns. In *Proceedings of ICDM-09*, page 764—769, 2009.

[2] C. Hashimoto, K. Torisawa, K. Kuroda, S. De Saeger, M. Murata, and J. Kazama. Large-scale verb entailment acquisition from the web. In *Proceedings of the EMNLP-09*, volume 3, page 1172—1181, 2009.

[3] J. Kazama and K. Torisawa. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. *Proceedings of ACL-08: HLT*, page 407—415, 2008.

[4] D. Lin and P. Pantel. Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(04):343—360, 2001.

[5] P. Malakasiotis and I. Androutsopoulos. Learning textual entailment using SVMs and string similarity measures. In *ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 2007.

[6] P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of COLING-ACL-06*, page 113—120, 2006.

[7] I. Varga, K. Ohtake, K. Torisawa, S. De Saeger, T. Misu, S. Matsuda, and J. Kazama. Similarity based language model construction for voice activated Open-Domain question answering. In *Proceedings of IJCNLP-11*, page 536—544, Chiang Mai, Thailand, November 2011.