Two-stage Method for Large-scale Acquisition of Contradiction Pattern Pairs using Entailment

Julien Kloetzer[*] **Stijn De Saeger**[†] **Kentaro Torisawa**[‡] **Chikara Hashimoto**[§]
Jong-Hoon Oh[¶] **Motoki Sano**[‖] **Kiyonori Ohtake**[**]
Information Analysis Laboratory,
National Institute of Information and Communications Technology (NICT), Kyoto, Japan
{[*]julien, [†]stijn, [‡]torisawa, [§]ch, [¶]rovellia, [‖]msano, [**]kiyonori.ohtake}@nict.go.jp

Abstract

In this paper we propose a two-stage method to acquire contradiction relations between typed lexico-syntactic patterns such as X_{drug} *prevents* $Y_{disease}$ and $Y_{disease}$ *caused by* X_{drug}. In the first stage, we train an SVM classifier to detect contradiction pattern pairs in a large web archive by exploiting the *excitation* polarity (Hashimoto et al., 2012) of the patterns. In the second stage, we enlarge the first stage classifier's training data with new contradiction pairs obtained by combining the output of the first stage's classifier and that of an *entailment* classifier. We acquired this way 750,000 typed Japanese contradiction pattern pairs with an estimated precision of 80%. We plan to release this resource to the NLP community.

1 Introduction

The ability to detect contradictory information in text has many practical applications. Among those, Murakami et al. (2009) pointed out that a contradiction recognition system can detect conflicts and anomalies in large bodies of texts and flag them to help users identify unreliable information. For example, many Japanese web pages claim that *agaricus prevents cancer*, where agaricus is a species of mushroom found in a variety of commercial products. Although this has been accepted by many Japanese people, by Googling keywords ”agaricus”, ”promotes” and ”cancer”, we can find pages claiming that ”agaricus promotes cancer”, some of which point to a study authorized by the Japanese Ministry of Health, Labour and Welfare[1] reporting that a commercial product containing agaricus promoted cancer. Obviously, the existence of these pages casts serious doubt on the ability of agaricus to prevent cancer and encourages readers to dig more about this subject.

The above example suggests that recognizing contradictory information can guide users to a *true* fact. Likewise, we believe that contradiction recognition is also useful when dealing with non-factual information that occupy most of our daily lives. For instance, there is a big controversy recently whether Japan should join an economic partnership agreement called the *Trans Pacific Partnership (TPP)*, and quite serious but contradictory claims are plentiful in the mass media and on the web, e.g., *TPP will wipe out Japan's agricultural businesses* and *TPP will strengthen Japan's agricultural businesses*. Neither of these are facts; they are *predictions* that can only be realized or disputed after the underlying decision-making is done: joining or refusing the TPP.

Furthermore, after reading documents including contradictory predictions, one should notice that each of them is supported by a *convincing* theory that has no obvious defect, e.g., “Exports of Japan's agricultural products will increase thanks to the TPP” or “A large amount of low-price agricultural products will be imported to Japan due to the TPP”. Even if one of these predictions may just happen to be true because of unexpected reasons such as minor fluctuations in the Japanese yen, we must survey such theories that support contradictory predictions, conduct balanced decision-making, and prepare counter measures for the expected problems after examining multiple viewpoints. Contradiction recognition should be useful to select documents to be surveyed.

[1] http://www.mhlw.go.jp/topics/bukyoku/iyaku/syoku-anzen/qa/060213-1.html

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 693–703,
Seattle, Washington, USA, 18-21 October 2013.

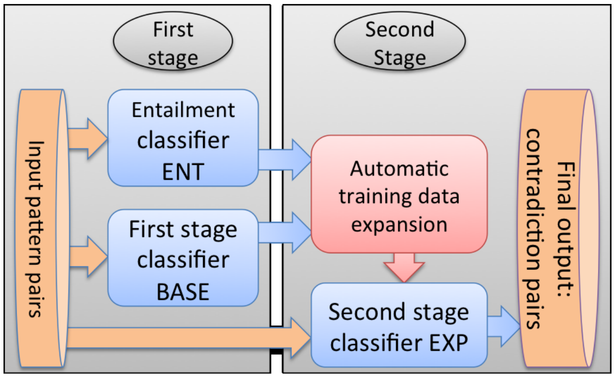
First stage
Second Stage
Input pattern pairs
Entailment classifier ENT
First stage classifier BASE
Automatic training data expansion
Second stage classifier EXP
Final output: contradiction pairs

Figure 1: Method workflow

We have developed a method for recognizing pairs of contradictory binary patterns such as ⟨“*X promotes Y*”, “*X prevents Y*”⟩ and ⟨“*X will wipe out Y*”, “*X will strengthen Y*”⟩. To solve the problem described above, we can easily develop a system that can find contradictory text fragments from the web like “*agaricus promotes cancer*” and “*agaricus prevents cancer*” from the discovered contradictory pattern pairs.

Our method is a two-stage procedure with three supervised classifiers (Fig. 1). In the first stage, we build a classifier **BASE** to recognize contradictions between binary patterns, and a classifier **ENT** to recognize entailment. In the second stage, we combine the contradiction pairs recognized by **BASE** and the entailment pairs recognized by **ENT** to expand **BASE**'s training data and train a new contradiction classifier, **EXP**. This expansion using entailment is one key idea of this work: we acquired 750,000 contradiction pairs with 80% precision using the expanded training data, more than doubling the 285,000 pairs acquired at the same precision level without expansion. We also demonstrate that this result is not trivial by showing that our method outperforms an alternative one based on *Integer Linear Programming* inspired by the successful entailment recognition method of Berant et al. (2011).

As another technical contribution of this work, we exploit the recently proposed semantic polarity of *excitation* (Hashimoto et al., 2012) to recognize contradictions between binary patterns. Hashimoto et al. (2012) previously showed that excitation polarities are useful to recognize contradictions between phrases that consist of a noun and a predicate, such as “promote cancer” and “prevent cancer”. While it is trivial to extend this framework to contradictions between *unary* patterns such as “promote X” and “prevent X” by replacing the common nouns in each pair with a variable, the information represented in unary patterns is often vague, and it is unlikely that a contradiction between unary patterns directly leads to the discovery of unreliable information to be flagged or to a meaningful survey of complex problems. As exemplified by the *agaricus* and *TPP* examples, contradictions between *binary* patterns that include two variables such as “X promotes Y” or “X will wipe out Y” are more useful than those between unary patterns. We also show that it is not trivial to recognize contradictions between binary patterns using contradictions between unary patterns.

Most works dealing with contradiction recognition up till now (Harabagiu et al., 2006; Bobrow et al., 2007; Kawahara et al., 2008; Kawahara et al., 2010; Ohki et al., 2011) focus on recognizing contradictions between full sentences or documents, not text fragments that match our relatively short patterns (survey in Section 5). We expect that the contradictory pattern pairs we acquired can be used as building blocks in such full-fledged contradiction recognition for full sentences or documents, similarly to antonym pairs in Harabagiu et al. (2006).

Also, we should emphasize that our method focuses on the most challenging part of contradiction recognition according to the classification of De Marneffe et al. (2008). Since we discard patterns with negations, an evident source of contradictions like ⟨“*X causes Y*”, “*X does not cause Y*”⟩, most of our output are non-trivial contradictions related to high-level semantic phenomena, e.g., contradiction pairs related to antonyms like ⟨“XがYを上げる”, “XがYを下げる”⟩ (⟨“*X increases Y*”, “*X decreases Y*”⟩), lexical contradictions like ⟨“XがYに勝つ”, “YがXに勝つ”⟩ (⟨“*X wins against Y*”, “*Y wins against X*”⟩), or contradictions due to common-sense knowledge like ⟨“XがYを安心させる”, “XがYを裏切る”⟩ (⟨“*X reassures Y*”, “*X betrays Y*”⟩). We believe acquiring such contradictions in a large scale is a valuable contribution.

The following is the outline of this paper. Section 2 details our target and our proposed method. Evaluation results are discussed in Section 3. Sec-

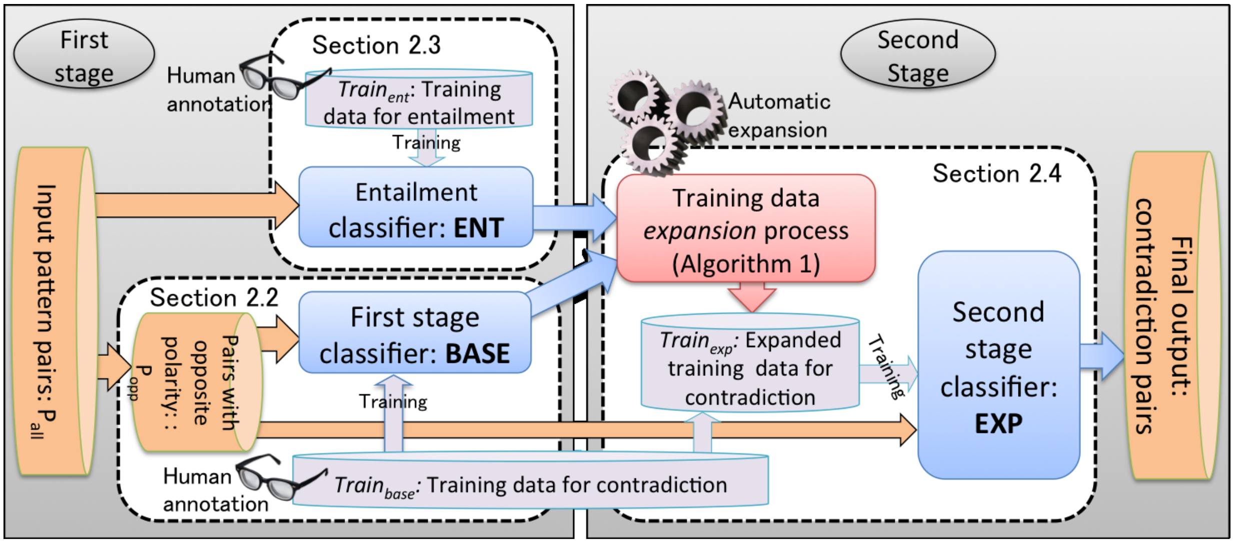
First stage
Human annotation
Section 2.3
Train_ent: Training data for entailment
Training
Entailment classifier: ENT
Input pattern pairs: P_all
Section 2.2
Pairs with opposite polarity: : P_opp
First stage classifier: BASE
Training
Human annotation
Train_base: Training data for contradiction
Second Stage
Automatic expansion
Section 2.4
Training data expansion process (Algorithm 1)
Train_exp: Expanded training data for contradiction
Training
Second stage classifier: EXP
Final output: contradiction pairs

Figure 2: Detailed data flow

tion 4 details our features set, and Section 5 related work. Section 6 provides a conclusion.

2 Proposed method

As showed in Figure 1, our method consists of three supervised classifiers. Classifiers **BASE** and **EXP** recognize contradiction relations between binary patterns, and **ENT** recognizes entailment relations between binary patterns. The contradiction pairs recognized by **BASE** and the entailment pairs recognized by **ENT** are combined to generate new contradiction pairs, part of which are then added to **BASE** training data to train the **EXP** classifier. Our final output is the set of all binary pattern pairs regarded as contradictions by **EXP**. Since the dependencies between these three classifiers, their distinct sets of training data, and the two data sets to be classified (we describe those in the two sections below) is a bit complex, we show a complete description of the whole process in Figure 2.

The key idea is in the scheme that expands the training data. Logically speaking, patterns p and r are contradictory if there exists a pattern q such that p entails q and q contradicts r. For example, since "*X causes Y*" entails "*X promotes Y*" and "*X promotes Y*" contradicts "*X prevents Y*", then "*X causes Y*" contradicts "*X prevents Y*". Hence, by combining entailment and contradiction pairs, we can obtain more contradiction pairs.

Following this property of contradiction relations, we collect a set of pattern pairs $\{\langle p, r \rangle\}$ for which there exists a pattern q such that **ENT** recognizes that p entails q and **BASE** recognizes that q contradicts r. Then we rank these pairs based on a novel scoring function called *Contradiction Derivation Precision (CDP)* and expand **BASE** training data by adding to it the top-ranked pairs according to CDP in order to train **EXP**. This ranking scheme selects highly accurate contradiction pairs and prevents errors caused by **BASE** and **ENT** from being propagated to **EXP**.

In the following, after defining the patterns for which we acquire contradiction relations, we describe **BASE**, **EXP**, **ENT**, and our expansion scheme.

2.1 Patterns

In this work, a binary pattern is a word sequence on the path of dependency relations connecting two nouns in a syntactic dependency tree, like "*X causes Y*", and we say a noun pair *co-occurs with* a pattern if the two nouns are connected by this pattern in the dependency tree of a sentence in the corpus.

We focus on *typed* binary patterns, which place semantic class restrictions on the noun pairs they co-occur with, e.g., "$Y_{organization}$ *is in* $X_{location}$". Subscripts *organization* and *location* indicate the semantic classes of the X and Y slots. Since typed patterns can distinguish between multiple senses of ambiguous patterns, they greatly reduce errors due to pattern ambiguity (De Saeger et al., 2009; Schoenmackers et al., 2010; Berant et al., 2011). We automatically induced semantic classes from our corpus using the EM-based noun clustering algorithm presented in Kazama and Torisawa (2008), and clustered one million nouns into 500 relatively clean semantic classes, including for example classes of *diseases* and of *chemical substances*.

The binary patterns and their co-occurring noun pairs were extracted from our corpus of 600 million Japanese web pages dependency parsed with KNP (Kurohashi and Nagao, 1994). We restricted our patterns to the most frequent 3.9 million patterns of the form "*X-[case particle] Y-[case particle] predicate*" such as "*X-ga Y-ni aru*" ("*X is in Y*") which do not contain any negation, number, symbol or punctuation character. Based on our observation that patterns in meaningful contradiction and entailment pairs tend to share many co-occurring noun pairs, we used as input to our classifiers the set P_{all} of 792 million pattern pairs for which both patterns share three co-occurring noun pairs.

2.2 BASE: First stage Classifier for Contradiction

Below, we detail BASE: its training data and input data to be classified, and some experimental results.

Our first stage classifier for contradictions, BASE, is an SVM that uses commonsensical surface and lexical resources based features, such as n-grams extracted from patterns, which will be detailed in Section 4. An important point to be stressed here is that we restricted the pattern pairs to be classified by BASE by exploiting their *excitation* polarity, a semantic orientation proposed by Hashimoto et al. (2012). Excitation characterizes unary patterns as *excitatory*, *inhibitory*, or *neutral*. *Excitatory* unary patterns, such as "*cause X*" or "*increase X*", entail that the function, effect, purpose, or role of their argument's referent is activated or enhanced, and *inhibitory* unary patterns, such as "*prevent X*" or "*X disappears*", entail that the function, effect, purpose, or role of their argument's referent is deactivated or suppressed. Neutral unary patterns like "*close to X*" are neither excitatory nor inhibitory.

We exploited *excitation* to restrict the input of BASE. Based on the result of Hashimoto et al. (2012) showing that two unary patterns with opposite polarity have a higher chance to be a contradiction, we extracted from set P_{all} the set P_{opp} of binary pattern pairs that contain unary patterns with opposite excitation polarities as sub-patterns. ⟨"*Y cause X*", "*Y prevent X*"⟩ is an example of such a pair since the unary sub-patterns "cause X" and "prevent X" are respectively excitatory and inhibitory. We used here 6,470 excitation unary patterns hand-labeled as either excitatory (4,882 patterns) or inhibitory (1,588 patterns). Set P_{opp} contains 8 million pattern pairs with roughly 38% *true* contradiction pairs, and is the input to BASE. We will show in experiments at the end of this section that this restriction is necessary to obtain good performance for BASE. We also tried to add the excitation polarities in BASE's feature set and classify P_{all}, but the performance was worse.

Training Data Another key feature of BASE is that it is distantly supervised. We did not use training samples that are directly manually annotated. Instead we automatically generated training data from a smaller set of (non-)contradiction unary pattern pairs. We first prepared a set of roughly 800 unary pattern pairs hand-labeled by three human annotators as contradictions (238 pairs) and non-contradictions (558 pairs) using majority vote. The inter-annotator agreement was 0.78 (Fleiss' kappa). Inspired by Hashimoto et al. (2012), we selected these unary pattern pairs among pairs with high distributional similarity, with and without restricting them to having opposite excitation polarity, such as to get a fair distribution of contradictions and non-contradictions.

We then extracted from set P_{all} all 256,000 pattern pairs containing a contradictory unary pattern pair, and all 5.2 million pattern pairs containing a non-contradictory unary pattern pair, which we respectively used as positive and negative training data (estimated 79% and 73% accuracy from 200 hand-labeled samples). Table 1 shows some examples.

The optimal composition of training data for BASE was determined according to preliminary experiments using our development set (1,000 manually labelled samples. See Section 3.1). We trained 20 different classifiers using from 6,250 to 50,000 positive samples (4 sets) and from 12,500 to 200,000 negative samples (5 sets), doubling the amounts in each step, for a total of 20 configurations. We could not try a larger training data due to long training time but we do not expect it to be a problem because the worst performance was observed with large train-

Table 1: Examples of training samples for **BASE** obtained from unary pattern pairs

Binary pattern pair (the unary pattern pair that extracted it is underlined)	Unary pattern pair label
Y も X が悪い (X is bad in Y too) - Y でも X が良い (X is good even in Y)	contradiction
Y も X に向かう (Y too heads toward X) - Y も X を出る (Y too comes out of X)	contradiction
X にY を 加える (add Y to X) - X をY に 入れる (insert X into Y)	non-contradiction
Y も X に来る (Y too comes to X) - Y とは X に行く (go to X with Y)	non-contradiction

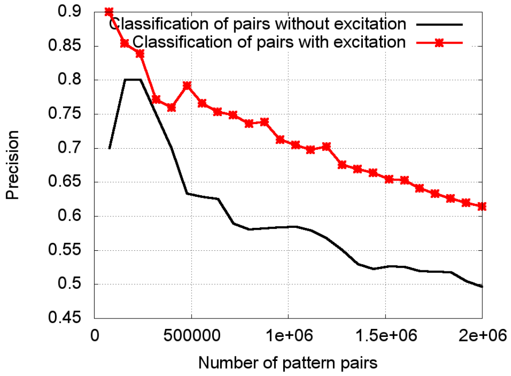

Figure 3: Effect of the restriction using excitation

ing data (25,000 positives and 200,000 negatives; the difference from the optimal setting was 2.3% in average precision). The optimal training data set, $Train_{base}$, consists of 12,500 positives and 100,000 negatives samples as described above and is the one we use in our experiments below and in Section 3.

Since **BASE** input for classification data is P_{opp} we also tried sampling $Train_{base}$ from P_{opp}. We obtained 56.27% average precision for our classifier **BASE**, and 52.99% when restricting the source of training data to pairs in P_{opp}. We believe that the difference lies in the size of the sets from which we sampled our training data: while there are 5.46 million binary pattern pairs in P_{all} with a hand-labeled unary pattern pair in P_{all}, there are only 237,000 pairs in P_{opp}. We believe this much smaller sample source lead to a lower performance because it included much less variations of the patterns.

To train **BASE** and other classifiers mentioned in this paper, we used the SVM tool *TinySVM*[2] with a polynomial kernel of degree 2, the setting which showed the best performance during our preliminary experiments.

Effect of Excitation Polarities We also empirically examined the effect of the restriction on the patterns using excitation polarities. We used our test set (2,000 manually annotated samples described in Section 3.1) and 250 manually annotated samples (majority vote from 3 annotators) from top ranked pairs of P_{all} to draw precision curves for **BASE** over the top 2 million binary pairs from both P_{opp} and P_{all}. In each case we assumed that pairs were distributed uniformly (i.e., with a constant interval) in the ranked list of pairs of P_{opp} and P_{all}, and computed precision accordingly. Since the pairs sets are reasonably large and were sampled randomly we thought this was a reasonable hypothesis. The precision over P_{opp} is higher than that over P_{all} with a large margin, suggesting that the restriction using excitation polarities is beneficial.

2.3 ENT: First stage Classifier for Entailment

ENT is an SVM classifier for entailment trained using 27,500 hand-annotated binary pattern pairs (set $Train_{ent}$, 45% of positive entailment pairs) created for some previous work (Kloetzer et al., 2013). It essentially uses the same feature set as that for **BASE** with the addition of several *distributional* similarity measures (see Section 4 below for more details). This classifier is given all pairs of P_{all} as input and scores each of them. For this study, we considered the 44.5 million pattern pairs with a positive SVM score as entailment pairs. Manual annotation of 200 random samples revealed that the precision of these pairs was 63% and that the top 7.1 million pairs had 80% precision (result interpolated from the top 16% of the annotated samples).

2.4 Second stage: Training Data Expansion and Classifier EXP

Below, we show how we combine **BASE**'s top output (hereafter C) and **ENT**'s top output (hereafter E) in the second stage of our method to expand $Train_{base}$ and train a new classifier, **EXP**.

The training data expansion process is based on the following logical constraint: if a pattern p entails a pattern q and pattern q contradicts a third pattern r, then p must contradict r. For example, because "*X*

[2] http://chasen.org/˜taku/software/TinySVM/

Table 2: Examples of triplets $\langle p, q,r \rangle$ where p entails q, q contradicts r, and hence p contradicts r

Pattern p	Pattern q	Pattern r	X/Y examples	$SVM\ Score(p, r)$	$CDP(p, r)$
YからXが消える X disappears from Y	YからXが無くなる X vanishes from Y	YがXに満ちる Y is full of X	怒り/眼 anger/eye	0.3	0.98
YにXを停止する stop X in Y	YにXを終える finish X in Y	YからXを始める start X in Y	4月/活動 April/activity	-0.3	0.61
XはYを示す X shows Y	XがYを持つ X have Y	XはYを失う X loses Y	チーム/自信 team/confidence	0.07	0.45

Algorithm 1 Training data expansion: C is the top 5% output of **BASE**, E is the top output of **ENT** (score > 0)

```
procedure EXPAND(C, E)
    Compute the set of expanded pairs $C' = \{\langle p, r \rangle \mid \exists q : \langle p, q \rangle \in E, \langle q, r \rangle \in C\}$.
    Rank the pairs in $C'$ using CDP.
    Add the $N$ top-ranked pairs in $C' \setminus C$ as new positive samples to $Train_{base}$.
    Remove incoherent negative training samples using negative cleaning.
end procedure
```

causes Y" (pattern p) entails "*X promotes Y*" (pattern q) and the latter contradicts "*X prevents Y*" (pattern r), we conclude that "*X causes Y*" (p) contradicts "*X prevents Y*" (r). We call the former contradiction $\langle q, r \rangle$ a *source* contradiction pair, and the later pair $\langle p, r \rangle$ an *expanded* contradiction pair. Based on this idea, we combine C and E to aggressively expand $Train_{base}$. This process is described in Algorithm 1, and Table 2 shows examples of triples $\langle p, q,r \rangle$ obtained in our experiments.

Expanding pairs from C and E compounds the errors made by **BASE** and **ENT**, hence it is crucial to select a highly precise subset of the expanded pairs. Taking the top pairs according to their SVM score would achieve this, but since **BASE** already handles correctly such pairs, they should not help much as new training data. We therefore propose a new scoring function for selecting highly precise expanded pairs: *Contradiction Derivation Precision* (CDP).

CDP was designed according to the following assumption: a source contradiction pair that derives *correct* expanded pairs with a high *precision* should be *reliable*. Probably, *all* the expanded pairs derived from such a reliable source pair will be *correct* and should be included in the new training data .

In our formulation of CDP, correctness of an expanded pair is judged according to the pair's SVM score using **BASE**. In other words, we regard an expanded pair that has an SVM score above some threshold α as a *true* contradiction. A source contradiction pair that derives *true* contradiction pairs with a high *precision* is regarded as a *reliable* source contradiction pair. CDP, which is defined over a expanded pairs, is the maximum precision among that of the source contradiction pairs that derive a given expanded pair.

We first define $CDPsub(q, r)$ over a source contradiction pair $\langle q, r \rangle$ as the ratio of expanded pairs obtained from $\langle q, r \rangle$ whose SVM score is above threshold α. This ratio corresponds to the *precision* of the expanded pairs derived from the source contradiction pair $\langle q, r \rangle$.

$$CDPsub(q, r) = \frac{|\{\langle p, r \rangle \in Ex(q, r) \mid Sc(p, r) > \alpha\}|}{|Ex(q, r)|}$$

Here $Ex(q, r)$ is the set of expanded pairs derived from a source pair $\langle q, r \rangle$, and Sc is the SVM score given by **BASE**. In our experiments, we set $\alpha = 0.46$ such that pattern pairs for which **BASE** gives a score over α corresponds to the top 5% of **BASE**'s output. $CDP(p, r)$ over an expanded pair is defined as follows, where $Source(p, r)$ is the set of source contradiction pairs that were derived into the expanded pair $\langle p, r \rangle$.

$$CDP(p, r) = max_{\langle q, r \rangle \in Source(p, r)} CDPsub(q, r)$$

We then expand the top 5% contradictions of **BASE**'s output (set C) and pattern pairs scored positively by **ENT** (set E), rank all expanded pairs not already in C according to CDP, and add the top N pairs with the highest CDP values as positives to $Train_{base}$ to train **EXP**. The value of N shall be determined empirically in later experiments using a development set. Note that, since $CDP(p, r)$ is independent of $\langle p, r \rangle$'s SVM score, even pairs that were assigned a negative score by **BASE** can become highly ranked by CDP (second triplet in Table 2) and be added to train **EXP**, hence we expect **EXP** to learn something new from these pairs.

Finally, after the addition of expanded pairs, we remove incoherent training samples. We propose to remove from the *negative* training samples of **EXP** any pattern pair that may conflict with the newly added positives; we call this step *negative cleaning*. Intuitively, since the content word pairs in a pattern pair should present some of the strongest evidence for determining the patterns (non-)contradiction status, we remove any negative sample that shares a content word pair with one of the added expanded pairs. The final training data for **EXP**, set $Train_{exp}$, consists of the following: **(1)** positive samples from $Train_{base}$, **(2)** (positive) expanded pairs, and **(3)** negative training samples from $Train_{base}$, cleaned using *negative cleaning*. We confirmed in our experiments that *negative cleaning* was necessary to train a strong **EXP** classifier (details omitted for reason of space).

After training **EXP** with $Train_{exp}$, we classify P_{opp} with **EXP** to produce the final output of the whole method. Note that while this expansion process can be re-iterated with **EXP**'s output, our experiments failed to show any improvement with subsequent iterations.

3 Evaluation

This section presents our experimental results. We describe first how we constructed test and development data, and then report comparison results between our method and others including **BASE** and an Integer Linear Programming-based (ILP) method.

3.1 Development and Test Data

We asked three human annotators to label 3,000 binary pattern pairs randomly sampled from P_{opp} as contradiction or non-contradiction to be used as development (1,000 pairs) and test (2,000 pairs) sets. We considered a pattern pair as a true contradiction relation if at least two out of the three annotators marked it as positive. The inter-rater agreement score (Fleiss Kappa) was 0.523, indicating moderate agreement (Landis and Koch, 1977). As a definition of contradiction, we used the notion of *incompatibility* (i.e., two statements are extremely unlikely to be simultaneously true) proposed by De Marneffe et al. (2008). We then say binary patterns such as "*X causes Y*" and "*X prevents Y*" are contradictory if the above definition holds for any noun pair that can instantiate the patterns' variables in the provided semantic class pair.

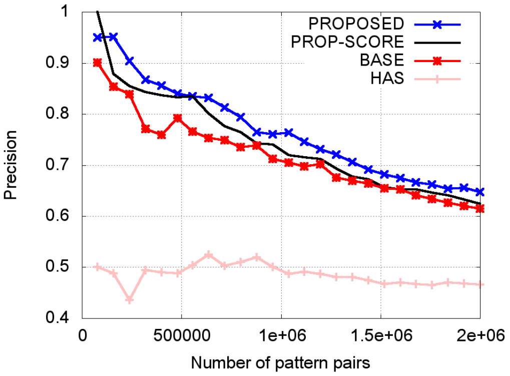

Figure 4: Precision of all the compared methods

Because our semantic classes are obtained by automatic clustering and have no meaningful labels, we followed Szpektor et al. (2007) and provided the annotators with three random noun pairs that cooccur with the patterns as a proxy for the class pair. The annotators marked a given pattern pair as positive if the contradiction relation between the patterns held for all three noun pairs presented.

3.2 Experimental Results

Here we show how our proposed method outperforms baseline methods. We compare the following four methods:

- **PROPOSED**: our proposed method. N, the number of newly added positive training samples during the training data expansion process, was set to 6,000 according to preliminary experiments using the development set. We tried 50 different values of N from 1,000 up to 50,000, adding 1,000 each time, and chose the N value giving the highest average precision against our development set (1,000 samples).
- **BASE**: our first stage classifier.
- **PROP-SCORE**: same as **PROPOSED** except for the use of **BASE**'s SVM score instead of CDP. N was set to 30,000 in the same way we set N for **PROPOSED**.
- **HAS**: an adaptation of the contradiction extraction method presented in Hashimoto et al.

(2012). For a binary pattern pair we first extracted its unary pattern pair with opposite polarity (or one at random in case there are two) and scored it based on our implementation of Hashimoto et al. (2012); the score is based on the distributional similarity between unary patterns and an excitation score obtained using a minimally supervised method based on the spin model. We then scored the binary pattern pair by the score of this unary pattern pair.

We ranked the pattern pairs of our test set (2,000 random pairs from set P_{opp}) based on the score produced by each method. For each tested method we assumed that pairs in the test set were distributed uniformly like explained in Section 2.2. The precision curves we obtained are shown in Figure 4.

PROPOSED clearly outperformed BASE and acquired around 750,000 contradiction pattern pairs with an estimated precision of 80%, out of which some examples are shown in Table 3. These pairs cover 26,941 content word pairs and reduce to 272,164 *untyped* pairs, showing that PROPOSED does not just acquire a handful of contradictions in many different class pairs. Also, when matching these pairs against an antonyms database (extracted from the dictionary of the morphical analyzer JUMAN) we found that only 100,886 of these pattern pairs contain an antonym pair, which means that most of the extracted pairs' contradictions are due to more complex phenomena than simple antonymy.

With the same precision, BASE and PROP-SCORE acquired only 285,000 pairs (covering 11,794 content word pairs) and 636,000 pairs respectively. This implies that our two-stage method can more than double the number of highly precise contradiction pairs we acquire as well as increasing their variety, and that ranking expanded pairs using our scoring function CDP is better than with SVM score, though even PROP-SCORE performs better than BASE in our setting. Finally, the poor performance of HAS suggests that extending the Hashimoto et al.'s framework to recognition of binary patterns is not a trivial task.

As to why adding only 6,000 top pairs ranked by CDP performs better than adding 30,000 pairs ranked by SVM score, the pattern pairs added in PROP-SCORE had high SVM scores given by BASE and as such are already handled nicely by BASE. Hence, we think the effect of adding a new sample from PROP-SCORE is smaller than that in PROPOSED, because in PROPOSED we add to the training data pattern pairs with both high and low (possibly negative) SVM scores.

Table 3: Examples of pairs acquired by PROPOSED: contradiction (label +) and non-contradiction (label -)

Lab.	Pattern pairs (with rank)	X/Y example
+	YでXが終わる - YよりXを開始する X finished Y - X started from Y Rank 228,039	販売/昨日 sale/yesterday
+	XがYに勝つ - YがXに勝つ X wins against Y - Y wins against X Rank: 258,068	日本/ベトナム Japan/Vietnam
-	XはYを失う - XにはYはある X lose Y - Have Y in X Rank 474,143	人/興味 people/interest
+	YにXを無くす - YにもXをもつ Lose X in Y - Have X in Y too Rank 522,534	自信/自分 confidence/ oneself
-	YはXまで落ちる - XにYを上げる Y falls down to X - raise Y to X Rank 538,901	9位/順位 9th/ranking
+	XにYが存在する - XからYを防ぐ Y exists in X - Keep Y out X Rank 620,430	中/ウイルス inside/virus
-	XからYを外す - XはYで答える Remove Y off X - X answer with Y Rank 652,530	僕/目 I (or me)/eyes
+	YをXから追い出す - XにYが残る Kick out Y from X - Y remains in X Rank 697,177	体/疲労 body/fatigue
+	YがXを安心させる - YがXがを裏切る X reassures Y - X betrays Y Rank: 749,916	僕/彼女 I/her

Finally, while the quality of the entailment pairs plays a very important role in the assumption that was the base of CDP, these results show that even a simple rule such as "Use entailment pairs with SVM score over 0 to expand contradictions before ranking them with CDP" is sufficient to make the method work. Though it may be possible to design a more complex CDP formula which takes entailment score into account, we did not explore this direction in this work.

Comparison with an ILP-based method Finally, we would like to compare our method with an ILP-based method. The interaction between contradiction and entailment that forms the basis for our expansion method has a natural interpretation as an optimization problem. We thus compared our method to the following ILP formulation of this interaction inspired by Berant et al. (2011), using our test set:

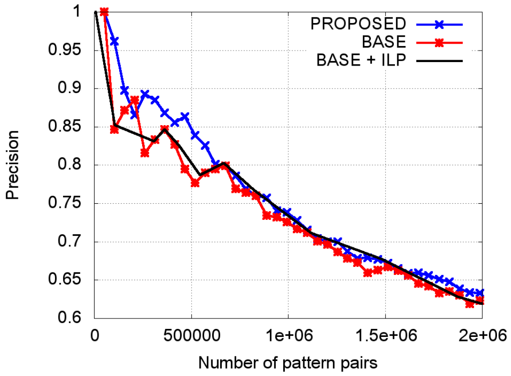

Figure 5: Comparison between **PROPOSED**, **BASE** and **BASE+ILP** on a restricted test set (1,306 samples)

(1) $G = argmax \sum_{p \neq q} (e(p,q) - \beta) * E_{pq} + (c(p,q) - \beta) * C_{pq}$

(2) s.t. $\forall_{p,q,r}\ E_{pq} + C_{qr} - C_{pr} \leq 1$

(3) $\forall_{p,q}\ E_{pq} + C_{pq} \leq 1$

(4) $\forall_{p,q}\ E_{pq} \in \{0,1\}$ **(5)** $\forall_{p,q}\ C_{pq} \in \{0,1\}$

The objective in Equation (1) is a sum over the weights of every pair of patterns $\langle p, q \rangle$, where E_{pq} indicates whether a pair $\langle p, q \rangle$ is an entailment pair (Equation (4)), and C_{pq} indicates whether it is a contradiction pair (Equation (5)). $e(p,q)$ and $c(p,q)$ are the score given respectively by **ENT** and **BASE**, and β is a prior defining the weight of a pair as neither entailment nor contradiction that shall be set before any experimentation. Equation (2) states the transitivity relation which is the basis of our expansion method. Finally, Equation (3) states that a given pattern pair cannot be a contradiction pair and an entailment pair at the same time. Since our patterns are class-dependent, we solved separate ILP instances for each semantic class pair.

We drew a precision curve for each of **BASE**, **PROPOSED** and **BASE+ILP**. To draw the curve for **BASE+ILP**, we incrementally raised the sample's non-contradiction non-entailment prior β (more details in Berant et al. (2011)). Because of the computational difficulty of ILP (NP-complete) and the size of our data, the computation for the ILP-based method ran out of memory on a 72GB machine for 116 class pairs out of the 1,031 that our test set covers. For this reason, we only used the 1,306 samples of the test set covered by the remaining 915 class pairs. We also measured the performance of **BASE** and **PROPOSED** on the same restricted test set.

Figure 5 shows that under these conditions the ILP-based method performance resembles **BASE** and is worse than **PROPOSED** on all data points. **PROPOSED** performs slightly worse in this setting compared to when classifying the whole of P_{opp}, but this only means that its performance is good for the 116 class pairs we ignored in this experiment. While this comparison is only made in a restricted setting, our expansion method still outperforms ILP and is clearly more scalable. The ILP results could be improved by adding more constraints (*contradiction is symmetric*, *entailment is transitive*), but this would also make the problem even more intractable in terms of computational costs.

4 Features

In this section we present the features used in our classifiers, which are mainly categorized into three: surface features (i.e., those reflecting the patterns' content itself), features based on external lexical resources, and distributional similarity based features; all features are listed in Table 4. **ENT** uses all the features while **BASE** and **EXP** use all except for the distributional similarity based ones. The optimality of the feature sets was confirmed through ablation tests using the development set (results omitted for the sake of space).

Since patterns with a contradiction or entailment relation are often superficially similar, for instance, in case structure or inflection, we use a number of *surface* features based on string similarity measures, extending the feature sets used by Malakasiotis and Androutsopoulos (2007) for entailment recognition. They include bag-of-words features such as n-grams and similarity scores concerning the bag-of-words such as their Euclidian distance.

To complement the surface features with knowledge about the content words, we used lexical databases including such as antonymy, synonymy, entailment, or allography. The presence of such word pairs is usually a good indicator of (non-)contradiction or (non-)entailment at the pattern level. More specifically, for any word pair $\langle wp, wq \rangle$ taken from a pattern pair $\langle p, q \rangle$ we mark the presence of $\langle wp, wq \rangle$ in each of the lexical resources as a binary feature. We used the Japanese lexical resources distributed by the *ALAGIN Forum*[3]: the verb entailment database (117,000 verb

[3] http://www.alagin.jp/

Table 4: Features summary, computed over a pair of patterns $\langle p, q \rangle$

surface	**Similarity measures:** common elements ratios, Dice coefficient, Jaccard and discounted Jaccard scores, Cosine, Euclidian, Manhattan, Levenshtein and Jaro distances; *computed over:* the patterns' 1-, 2- and 3-grams sets of: characters, morphemes, their stems & POS; content words and stems
	binary feature for each of the patterns' subtrees, 1- and 2-grams ; patterns' lengths and length ratios
lex.r.	entries in databases of verb entailments and non-entailments, synonyms, antonyms, allographs ; *checked over:* pairs of content words, pairs of content word stems, same for the reverse pattern pair $\langle q, p \rangle$
dis.s.	**Distributional similarity measures:** Common elements ratios, Jaccard and discounted Jaccard scores, sets and sets intersection cardinality, DIRT (Lin and Pantel, 2001), Weeds (Weeds and Weir, 2003) and Hashimoto (Hashimoto et al., 2009) scores; *computed over*: patterns' co-occurring noun pairs, POS tags of those, nouns co-occurring in each variable slot, nouns co-occurring with each unary sub-patterns
other	binary feature for each semantic class pair and individual semantic classes
	patterns frequency rank in the given semantic class pair

pairs; Alagin ID A-2), the databases of synonyms, antonyms and meronyms (respectively 111,000, 5000 and 2500 pairs; Alagin ID A-9), and the allographic word database (2.7 million pairs; Alagin ID A-7). We also used the information concerning allographic words in the dictionary of the morphological analyzer JUMAN[4].

Distributional similarity values between patterns are based on the idea that patterns that appear in similar contexts tend to have similar meanings and as such are useful to recognize entailment (Lin and Pantel, 2001). We computed as features several distributional similarity measures on the sets of each pattern's co-occurring noun pairs and their POS tags, of nouns co-occurring in each variable slot, and with each of the pattern's unary sub-patterns.

We also added a few more uncategorizable features. See Table 4 for more details.

5 Related Work

A number of previous work dealt with the recognition of contradictions between sentences. Harabagiu et al. (2006) proposed a contradiction detection method that focuses on negation, antonymy and some discourse information. Kawahara et al. (2010) also used negations and antonyms to extract contrastive/contradictory statements from the web to present users with a bird ’ s-eye view of statements about a given topic. Bobrow et al. (2007) showed a method using logical forms with relatively precise results. Ohki et al. (2011) proposed a method to recognize *confinment*, a novel semantic relation related to both entailment and contradiction. While we do not deal ourselves directly with sentences, we expect that the binary pattern pairs we acquire can play a role similar to that of basic linguistic resources such as antonyms and negations in these works. Closer to our work, Ritter et al. (2008) presented a method for detecting contradictions between *functional* relations like "*X was born in Y*", but these constitute only a part of the semantic relations expressed by the binary patterns we deal with in this paper.

Other works analyzed contradictions from linguistic/semantic viewpoints. Voorhees (2008) analyzed the contradiction recognition-task of the RTE3 contest. Magnini and Cabrio (2010) examined relations between contradictions and textual entailment samples. De Marneffe et al. (2008) presented a typology of contradictions, and showed that contradictions can arise from a multitude of phenomena. They showed contradictions based on lexical or world knowledge are challenging and require a high-level understanding of language and/or the world. As stated in the introduction, these are the types of contradictions our method focuses on.

6 Conclusion

This paper showed how to acquire a large number of contradiction pairs between lexico-syntactic binary patterns by exploiting **(1)** the interaction between contradiction and entailment, and **(2)** *excitation* polarities. In the end, we could acquire 750,000 typed contradiction pattern pairs with an estimated 80% precision. The resulting contradiction pairs covered ones deeply related to world knowledge such as the pair ⟨"*X reassures Y*", "*X betrays Y*"⟩. We expect our work to lead to a high level analysis of textual information, such as flagging unreliable information or identifying important documents to be surveyed for understanding complex social problems. We plan to release the data we acquired to the NLP community through the *ALAGIN Forum*[5].

[4] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

[5] http://www.alagin.jp/

References

J. Berant, I. Dagan, and J. Goldberger. 2011. Global learning of typed entailment rules. In *Proceedings of ACL 2011*, pages 610–619.

D. G. Bobrow, C. Condoravdi, R. Crouch, V. De Paiva, L. Karttunen, T. H. King, R. Nairn, L. Price, and A. Zaenen. 2007. Precision-focused textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, page 16—21.

M.-C. De Marneffe, A. N. Rafferty, and C. D. Manning. 2008. Finding contradictions in text. *Proceedings of ACL 2008*, page 1039—1047.

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

S.M. Harabagiu, A. Hickl, and V.F. Lacatusu. 2006. Negation, contrast and contradiction in text processing. In *Proceedings of AAAI 2006*, pages 755–762.

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

C. Hashimoto, K. Torisawa, S. De Saeger, J.-H. Oh, and J. Kazama. 2012. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of EMNLP 2012*.

D. Kawahara, S. Kurohashi, and K. Inui. 2008. Grasping major statements and their contradictions toward information credibility analysis of web contents. In *Proceedings of WI-IAT 2008*, volume 1, page 393—397.

D. Kawahara, K. Inui, and S. Kurohashi. 2010. Identifying contradictory and contrastive relations between statements to outline web information on a given topic. In *Proceedings of COLING 2010*, page 534—542.

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

J. Kloetzer, S. De Saeger, K. Torisawa, M. Sano, C. Hashimoto, and J. Gotoh. 2013. Large-scale acquisition of entailment pattern pairs. In *Information Processing Society of Japan (IPSJ) Kansai-Branch Convention*.

S. Kurohashi and M. Nagao. 1994. KN parser: Japanese dependency/case structure analyzer. In *Proceedings of the Workshop on Sharable Natural Language Resources*, page 48—55.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, page 159—174.

D. Lin and P. Pantel. 2001. Dirt - discovery of inference rules from text. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 323–328.

B. Magnini and E. Cabrio. 2010. Contradiction-focused qualitative evaluation of textual entailment. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, page 86—94.

P. Malakasiotis and I. Androutsopoulos. 2007. Learning textual entailment using SVMs and string similarity measures. In *Proceedings of the ACL- PASCAL Workshop on Textual Entailment and Paraphrasing*, page 42—47.

K. Murakami, E. Nichols, S. Matsuyoshi, A. Sumida, S. Masuda, K. Inui, and Y. Matumoto. 2009. Statement map: assisting information crediblity analysis by visualizing arguments. In *Proceedings of the 3rd workshop on Information credibility on the web*, page 43—50. ACM.

M. Ohki, S. Matsuyoshi, J. Mizuno, K. Inui, E. Nichols, K. Murakami, S. Masuda, and Y. Matsumoto. 2011. Recognizing confinement in web texts. In *the Proceedings of the Ninth International Conference on Computational Semantics*, page 215—224.

A. Ritter, D. Downey, S. Soderland, and O. Etzioni. 2008. It's a contradiction—no, it's not: a case study using functional relations. In *Proceedings of EMNLP 2008*, pages 11–20.

S. Schoenmackers, O. Etzioni, D. S Weld, and J. Davis. 2010. Learning first-order horn clauses from web text. In *Proceedings of EMNLP 2010*, page 1088—1098.

I. Szpektor, E. Shnarch, and I. Dagan. 2007. Instance-based evaluation of entailment rule acquisition. In *Proceedings of ACL 2007*, volume 45, page 456—463.

E. M. Voorhees. 2008. Contradictions and justifications: Extensions to the textual entailment task. In *Proceedings of ACL 2008*, page 63—71.

J. Weeds and D. Weir. 2003. A general framework for distributional similarity. In *Proceedings of EMNLP 2003*, page 81—88. Association for Computational Linguistics.