UB.dmirg: Learning Textual Entailment Relationships Using Lexical Semantic Features

Bahadorreza Ofoghi
Centre for Informatics & Applied Optimization, University of Ballarat
Victoria 3350, Australia
b.ofoghi@ballarat.edu.au

John Yearwood
Centre for Informatics & Applied Optimization, University of Ballarat
Victoria 3350, Australia
j.yearwood@ballarat.edu.au

Abstract

This paper describes our *Recognizing Textual Entailment* (RTE) system developed at University of Ballarat, Australia for participation in the Text Analysis Conference RTE 2010 competition. This year, we participated in the *Main* task and used a machine learning approach for learning textual entailment relationships using parse-free lexical semantic features. For this, we employed FrameNet and WordNet resources to extract event-based and semantic features from both hypotheses and texts. Our system also used the longest common substring of lemmas when learning the entailment relationships.

1 Introduction

Recognizing Textual Entailment (RTE) is formally described as automatically recognizing the relationship between a *hypothesis* and a *text*. The hypothesis (H) is a succinct piece of text and the text (T) includes a few sentences the meaning of which may or may not entail the truth/falsity of H. If the truth of H can be inferred from the evidence in T, then the relationship is denoted by $T \rightarrow H$.

With this view of RTE, a number of approaches to RTE have been developed in recent years. Systems that make use of morphological and lexical variations (Braz et al., 2005; Pazienza et al., 2005; Rodrigo et al., 2008), classical or plausible logic (Akhmatova and Molla, 2006; Tatu and Moldovan, 2005; Clark and Harrison, 2008), syntactic dependency trees (Lin and Pantel, 2001; Kouylekov and Magnini, 2005; Yatbaz, 2008), paraphrase detection (Bosma and Callison-Burch, 2006), and semantic roles (Braz et al., 2005) are among many existing RTE systems. Some other systems take a machine learning approach to learn entailment relationships (Corley and Mihalcea, 2005; Hickl et al., 2006; Mac Cartney et al., 2006; Zanzotto and Moschitti, 2006; Zanzotto et al., 2009; Ageno et al., 2008).

In 2010, the RTE challenge, conducted by the RTE Organizing Committee of the Text Analysis Conference (TAC), consisted of a *Main* task and a *Novelty Detection* task. The main task was concerned with the identification of all sentences that entail an hypothesis among a set of sentences retrieved by Lucene from a text corpus. The novelty detection task included the judgement of whether the information in a hypothesis was novel with respect to the information contained in the text corpus[1].

We participated in the RTE 2010 main task and our system followed the machine learning-based RTE strand. Our system extracted lexical and semantic features from texts. For this, we employed FrameNet (Baker et al., 1998) and WordNet (Miller et al., 1990) resources to extract event-based and semantic features from the lemmas of the hypotheses and texts. The

[1] See RTE-6 tasks guidelines available at http://www.nist.gov/tac/2010/RTE/RTE6_Main_NoveltyDetection_Task_Guidelines.pdf for more details.

longest common substring of lemmas was also used as a lexical feature.

In the next section, we describe the general structure of our RTE system participated in the TAC RTE 2010 competition. This is then followed by introducing the different textual features that our system used for learning the entailment relationships. We will then report the results of our system evaluated by TAC and finally conclude the paper.

2 UB.dmirg 2010 system overview

In the main task of the RTE 2010 challenge (i.e., recognizing textual entailment within a corpus), for each hypothesis, a set of candidate entailing sentences were retrieved by Lucene from the corpus. RTE systems were then required to identify all the sentences that entail a given hypothesis among the candidate sentences. In order to participate in this task, we:

- Developed a system to extract features from hypotheses and sentences (known as texts),
- Used a ready machine learning system to classify the entailment relationships between hypotheses and texts,
- Developed a system, given the class labels for each pair and their feature vectors, to align the two and put the feature vectors and their class labels into a repository, and
- Developed a system, given the class labels and feature vectors, to generate appropriate XML output according to the requirements of TAC for evaluation files.

Figure 1 shows a more fine-grained schematic overview of our work in the RTE 2010 main task. Given a T/H pair, the analysis started with lemmatizing both the hypothesis and text. The lemmas were then feed to our feature extractor program that used FrameNet and WordNet lexical resources to extract lexical, semantic, and event-based features and generate a feature vector for the pair. The feature vector was then added to a feature vector repository. The machine learning-based classification system received the feature vectors and in the training session, it generated classifier model. In the test session, however, the learning classification system produced class labels for the test set which were then aligned with the feature vectors of the pairs. The aligned repository was then processed by our other program to generate the final evaluation file in the XML format required by TAC.

2.1 Lemmatizer

Prior to feature extraction, all text and hypothesis terms were lemmatized using the *TreeTagger* lemmatizer (Schmid, 1994). This step was necessary in our system in order to normalize the terms in both texts and hypotheses and make it possible for the system to find that, for instance, both terms "*collapsed*" and "*collapsing*" are of the same root "*collapse*" and can be interpreted as an exact term match. In this year's competition, we did not remove stop words after lemmatization. The main reason for this was to maintain the structure of the sentences especially for when the longest common substring was calculated.

2.2 Feature extractor

The feature extractor function utilized the two aforementioned lexical resources, FrameNet and WordNet and extracts a number of lexical, semantic, and parse-free event-based features from both texts and hypotheses and generated a feature vector for each T/H pair. The feature vectors included the hypothesis identifier, the hypothesis topic identifier, the sentence (text) document (in the corpus) identifier, the sentence identifier, the lexical, semantic, and event-based feature values, and the evaluation result (class label).

The lexical features that we used were the total number of exact terms that matched between the text and hypothesis and also the longest common subsequence (LCS) of text and hypothesis lemmas. These features are among *similarity-based* features explained in (Burchardt et al., 2009).

The semantic features that we utilized were extracted by using WordNet lexical ontology. These features included:

– **Synonyms**: The total number of synonym terms that matched between the text and hy-

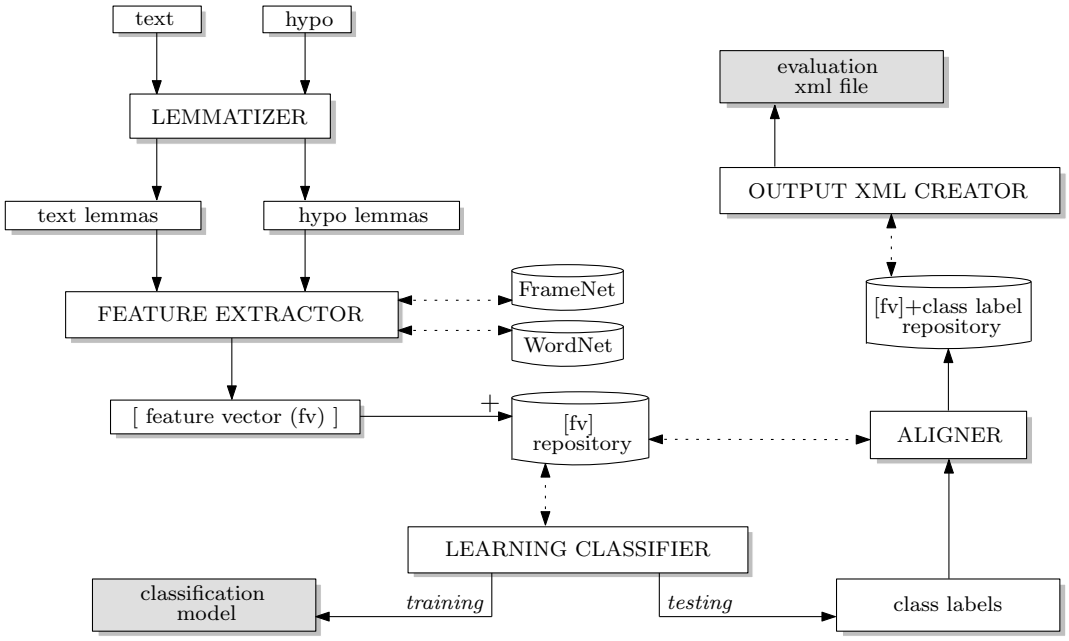

Figure 1: The schematic overview of UB.dmirg 2010 RTE system

pothesis term sets. This feature can overcome lexical paraphrasing.

– **Hypernyms/Hyponyms**: The total number of hypernyms and hyponyms that matched between the text and hypothesis term sets. The analysis of hypernym and hyponym matching took a *directional* approach. The procedure is formulated in Equation 1 where $s_{h/t}$ is the set of hypernyms or hyponyms for the hypothesis/text and $n_{h/t}$ represents the total number of terms t_i in the hypothesis/text. Hypernyms and hyponyms were extracted up to three links in WordNet. The idea behind using this directional approach was that WordNet hyponyms entail WordNet hypernyms e.g. "*female person*" lexically entails "*person*" but not vice versa.

This feature was designated to overcome the problem related to texts and hypotheses formulating concepts at different levels of conceptual abstraction. For instance, using this feature, it is possible to recognize the entailment relationship between "*Jack was in a European country last year.*" and "*Jack was in France last year.*" since "France" is a "European country" according to WordNet.

$$hyper/hypo\_score(h,t) = |s_h \cap s_t|$$
$$s_h = \bigcup_{i=1}^{n_h} hypernyms(t_i),\ s_t = \bigcup_{i=1}^{n_t} hyponyms(t_i) \quad (1)$$

– **Antonyms**: The antonym score was calculated using Equation 2 where $s_{h/t}$ is the set of exact terms or antonyms for the hypothesis/text term t_i. A similar attribute has been used as a trigger-based feature in (Burchardt et al., 2009). This feature can capture indications of contradiction or no-entailment relationships.

$$ant\_score(h,t) = |s_h \cap s_t|$$
$$s_h = \bigcup_{i=1}^{n_h} t_i,\ s_t = \bigcup_{i=1}^{n_t} antonyms(t_i) \quad (2)$$

– **Antonyms/Hyponyms**: This feature was measured using Equation 3 where $s_{h/t}$ is the set of hyponyms or antonyms for the hypothesis/text term t_i. This feature extended the last feature (antonyms) by looking for the occurrences of the antonyms of more specific terms of hypotheses in texts.

$$ant/hypo\_score(h,t) = |s_h \cap s_t|$$
$$s_h = \bigcup_{i=1}^{n_h} hyponyms(t_i),\ s_t = \bigcup_{i=1}^{n_t} antonyms(t_i) \quad (3)$$

We made use of FrameNet to extract two types of event-based features, namely *ebf* and *inter_ebf*. The ebf feature is the total number of FrameNet frames that are evoked both by text and hypothesis terms. To measure this, we used Equation 4 where the set of all frames that contain each hypothesis/text term was created by a term look-up procedure in FrameNet XML database. The union set of all framesets for all hypothesis/text terms was then created. The cardinality of the intersection of the two union sets was the score assigned to ebf. This method did not rely on any shallow semantic parser and/or word sense disambiguation procedure to evoke FrameNet frames; instead, was only based on fetching the names of frames that contain certain terms.

Using ebf, our RTE system can relate a hypothesis and a text that share semantics at the level of an event or state. This type of *scenario-based* similarity may not be captured using other types of lexical resources.

$$ebf\_score(h,t) = |s_h \cap s_t|$$
$$s_{h/t} = \bigcup_{i=1}^{n_{h/t}} frameset(t_i), frameset(t_i) = \bigcup_j frame_j$$
$$\{\exists t \in frame_j(termset) \mid t_i = t \wedge frame_j \in FN\_frames\} \quad (4)$$

FrameNet frames are inter-related using a number of frame-to-frame relations explained in detail in (Ruppenhofer et al., 2005). We used the *inheritance*, *subframe*, *using*, *inchoative-of*, *causative-of*, *precedes*, and *perspective-on* relations to extract one feature per relation type that represents another level of event/state-based textual similarity. We refer to this set of features as inter_ebf. For each type of frame-to-frame relation, we extracted FrameNet frames immediately inter-related to those frames evoked for each hypothesis/text term. A similar approach to the calculation of ebf was used to measure the overlap between inter-related frames evoked. Using inter_ebf, the system is able to recognize whether a hypothesis formulates a scenario that is, for instance, part of a big picture scenario or event (by measuring the overlap between inter-related frames according to the *inheritance* relationship).

In general, although our features were evoked on a term-basis procedure, the nature of the features and linguistic resources that we have used ensured that compositional semantics were indirectly taken into consideration.

2.3 Learning classifiers

For classifying the relationships between each T/H pair, we made use of Waikato Environment for Knowledge Analysis (WEKA)[2]. From a wide range of classification systems in WEKA, we used three classifiers: i) K-Nearest Neighbor classifier, ii) Random Forest classifier, and iii) Bayesian Network classifier.

Each classifier was first trained using the training data (the development set) and then used for generating class labels for each T/H pair in the test set. Since WEKA requires class labels for both training and test sets, in the training sessions, we used the gold standard class labels produced by TAC, whereas in the test sessions, we pre-produced dummy class labels for the feature vectors. These dummy class labels where not used for any other purposes than getting WEKA to run the classifiers in the test sessions.

2.4 Aligner and output xml generator

After each test session using each classifier was completed, the class labels were aligned with the feature vectors using our aligner system. This system merged each feature vector generated for each T/H pair with its class label predicted by the classifier. This was carried out by a sequential look-up in both class label file and feature vector repository.

Our other system was then employed to create the output XML file[3]. For each hypothesis

[2] http://www.cs.waikato.ac.nz/ml/weka/

[3] The structure of the XML file can be found

in each topic in the test set, the system looked for all of the sentences (texts) that entailed that given hypothesis (the class label was "*yes*"). It then added one entry, in the appropriate form required by TAC, to the hypothesis node in the output evaluation XML file for each sentence labels "*yes*".

The key to successful alignment of feature vectors and class labels as well as successfully generating the output XML file was the use of appropriate feature values that identified the topic, hypothesis, sentence, and document objects (see section 2.2).

3 Main system evaluation

3.1 Data

The main task of the RTE 2010 challenge consisted of a development set and a test set. Both the development set and the test set included 10 topics. Each topic consisted of a number of T/H pairs to be judged by RTE systems. Table 1 shows the statistics of these two datasets.

We used the development set for training purposes in the machine learning-based classifiers and the test set only for producing the class labels for the feature vectors of the T/H pairs.

3.2 System settings

Our UB.dmirg 2010 submissions included three different runs with the three machine learning-based classifiers introduced in section 2.3. Table 2 shows the settings of each run.

3.3 Results

Our three submissions to the main task were evaluated by TAC based on micro-averaged and macro- averaged precision, recall and f-measure as well as precision, recall and f-measure for each topic in the test set. Table 3 shows the micro-averaged and macro-averaged per topic measures of our runs. From these results, it can be see that the UB.dmirg3 run had the highest f-measure in terms of both micro-averaged and macro-averaged per topic measures. In fact, UB.dmirg3 had high recall measures (i.e. 48.68 and 49.79) for the two types of measures, although its precision values were comparable to those of the UB.dmirg1 and UB.dmirg2 runs.

According to TAC, 18 teams participated in the main task of the 2010 RTE challenge and submitted a total of 48 runs. The statistics over all of the runs showed micro-averaged f-measures 48.01, 33.72, and 11.60 as high, medium, and low bands. Compared to these statistics, the results of our three runs relied between the low and medium bands.

More detailed results, based on the precision, recall, and f-measure values of each run for each topic, are shown in Table 4.

4 Ablation tests

This year, as mentioned in section 2, our RTE system used two lexical resources, FrameNet and WordNet. To measure the contribution of each resource in our system, we ran two ablation tests for each run:

- Ablation test 1 (per run): we removed only the lexical and semantic features, extracted from T/H pairs using WordNet, from pair feature vectors and ran the machine learning-based classifier.
- Ablation test 2 (per run): we removed only the event-based features, extracted from T/H pairs using FrameNet, from pair feature vectors and ran the machine learning-based classifier.

Since one requirement, enforced by TAC, for the ablation tests was to remove only one component/resource from the original runs, we did not submit any run without both FrameNet and WordNet resources.

The results of the ablation tests shown in Table 5 do not reflect what we exactly expected. When removing either WordNet or FrameNet-based features from the learning process of the classifiers, we would have expected lower classification measures; however, in many cases, the evaluation measures show an increase without the features extracted using one of these lexical resources. The most consistent results were obtained by UB.dmirg3 ablation tests, where

at http://www.nist.gov/tac/2010/RTE/RTE6_Main_NoveltyDetection_Task_Guidelines.pdf

Table 1: The RTE 2010 main task development and test sets statistics reported by the TAC

Dataset	Topic	# of hypotheses	# of pairs to be judged
Development set	Topic 909 (IRA)	21	1169
	Topic 912 (Cindy Sheehan)	20	1928
	Topic 914 (Egypt attack)	13	944
	Topic 916 (US journalist abducted in Iraq)	21	1419
	Topic 922 (Patriot Act)	22	1457
	Topic 924 (Vioxx)	18	1754
	Topic 927 (Peter Jennings)	27	2700
	Topic 929 (Hurricane Rita)	19	616
	Topic 937 (Dick Cheney)	23	2134
	Topic 938 (WTC Memorial)	27	1834
Test set	Topic 901 (Kashmir)	27	2205
	Topic 902 (Morning-after pill)	26	1655
	Topic 907 (Torture and prison abuses)	27	1845
	Topic 913 (Ten Commandaments monuments)	20	1816
	Topic 918 (Betty Friedan)	23	2098
	Topic 928 (Tyco Trial)	28	2055
	Topic 931 (Abdul Qadeer Khan)	19	1900
	Topic 936 (Justice OConnor)	17	1340
	Topic 939 (LA train wreck)	27	2472
	Topic 943 (Air India bombings)	29	2586

Table 2: UB.dmirg 2010 system settings for each submitted run

Run	Classifier	Settings
UB.dmirg1	K-Nearest Neighbor	K=1 Distance weighting = none Search algorithm = linear NN search
UB.dmirg2	Random Forest	# of trees = 10 Seed = 1 Max depth = 0
UB.dmirg3	Bayesian Network	Estimator = simple estimator Search algorithm = K2

in both UB.dmirg3_abl-1 and UB.dmirg3_abl-2, only recall values decreased after removing one lexical resource. Both of the ablation tests of UB.dmirg2 show lower precision values compared to the main run whereas the ablation tests of UB.dmirg1 show the most inconsistent patterns of changes in classification measures.

In general, the nature of the ablation tests (where one component could be removed at a time), did not allow for a more fine-grained analysis of the impact of different individual features, but that of the whole lexical resource. Using more sophisticated methods, such as machine learning-based feature selection methods, one may be able to draw more insightful conclusions as to which features may more effectively be used for this type of RTE. A recent such study can be found in (Ofoghi and Yearwood, 2010).

5 Concluding remarks

A machine learning-based recognizing textual entailment system participated in the Text Analysis Conference (TAC) 2010 was introduced in this paper. The system, called UB.dmirg 2010, made use of a number of lexical, semantic, and event-based features extracted from two lexical resources, WordNet and FrameNet.

Table 3: Micro-averaged and macro-averaged per topic measures of UB.dmirg 2010 system reported by TAC

Run	Micro-averaged			Macro-averaged per topic		
	precision	recall	f-measure	precision	recall	f-measure
UB.dmirg1	12.22	13.44	12.80	12.58	13.42	12.99
UB.dmirg2	18.58	8.89	12.03	19.91	9.05	12.44
UB.dmirg3	11.79	48.68	18.98	12.48	49.79	19.96

The system did not make use of any semantic parser when extracting event-based features using FrameNet.

The UB.dmirg 2010 participation in TAC 2010 included three runs in the main task. We used three learning classifiers, K-Nearest Neighbor classifier, Random Forest classifier, and Bayesian Network classifier in the three runs. The results of our main runs, reported by TAC, showed that UB.dmirg's results were positioned between the low and medium bands of the results of all TAC 2010 participant systems/runs.

We submitted two ablation tests per run where in the first test, WordNet and in the second test, FrameNet was removed from the task. This meant that the classifiers of our system did not make use of the features extracted from WordNet and FrameNet in the two tests per run, respectively. The results of the ablation tests were not consistent with what we expected to see. In many cases, removing the features of one lexical resource did not result in any decrease in the classification measures of our system.

References

A. Ageno, D. Farwell, D. Ferres, F. Cruz, and H. Rodriguez. 2008. TALP at TAC 2008: A semantic approach to recognizing textual entailment. In *Proceedings of the Fourth PASCAL Challenges Workshop on Recognizing Textual Entailment*. Gaithersburg, USA.

A. Burchardt, M. Pennacchiotti, S. Thater, and M. Pinkal. 2009. Assessing the impact of frame semantics on textual entailment. *Natural Language Engineering*, 15(4):527–550.

A. Hickl, J. Bensley, J. Williams, K. Roberts, B. Rink, and Y. Shi. 2006. Recognizing textual entailment with LCCs GROUNDHOG system. In *Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment*, 80–85. Venice, Italy.

Alvaro Rodrigo, Anselmo Penas, and Felisa Verdejo. 2008. Towards an entity-based recognition of textual entailment. In *Proceedings of the Fourth PASCAL Challenges Workshop on Recognizing Textual Entailment*. Gaithersburg, Maryland, USA.

B. Mac Cartney, T. Grenager, M-C. de Marneffe, D. Cer, and C. D. Manning. 2006. Learning to recognize features of valid textual entailments. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. New York City.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING*, 86–90. Universite de Montreal, Montreal, Quebec, Canada.

C. Corley and R. Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. Ann Arbor, MI.

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*, 323–328. San Francisco, California, USA.

Elena Akhmatova and Diego Molla. 2006. Recognizing textual entailment via atomic propositions. In *Proceedings of the Machine Learning Challenges Workshop (MLCW)*, 385–403. Southampton, UK.

F. M. Zanzotto and A. Moschitti. 2006. Automatic learning of textual entailments with cross-pair similarities. In *Proceedings of the 21st COLING and 44th ACL*. Sydney, Australia.

F. M. Zanzotto, M. Pennacchiotti, and A. Moschitti. 2009. A machine learning approach to textual entailment recognition. *Natural Language Engineering*, 15(4):551–582.

George A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Table 4: Per topic measures of UB.dmirg 2010 system reported by TAC

Run	Topic	Precision	Recall	F-measure
UB.dmirg1	D0901	11.93	10.92	11.40
	D0902	8.51	4.30	5.71
	D0907	6.67	10.81	8.25
	D0913	8.62	18.29	11.72
	D0918	17.65	25.53	20.87
	D0928	15.53	16.16	15.84
	D0931	20.34	11.21	14.46
	D0936	17.65	13.24	15.13
	D0939	6.56	6.96	6.75
	D0943	12.36	16.79	14.24
UB.dmirg2	D0901	16.22	5.04	7.69
	D0902	22.22	6.45	10.00
	D0907	14.29	8.11	10.34
	D0913	11.39	10.98	11.18
	D0918	20.83	15.96	18.07
	D0928	16.98	9.09	11.84
	D0931	16.67	6.54	9.40
	D0936	38.10	11.76	17.98
	D0939	9.80	4.35	6.02
	D0943	32.65	12.21	17.78
UB.dmirg3	D0901	21.19	42.02	28.17
	D0902	10.20	50.54	16.97
	D0907	7.64	59.46	13.54
	D0913	8.28	78.05	14.97
	D0918	10.66	27.66	15.38
	D0928	17.71	51.52	26.36
	D0931	13.33	37.38	19.66
	D0936	9.52	50.00	16.00
	D0939	13.04	41.74	19.88
	D0943	13.27	59.54	21.70

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the Conference on New Methods in Language Processing*. Manchester, UK.

Josef Ruppenhofer and Michael Ellsworth and Miriam R.L. Petruck and Christopher R. Johnson. 2005. FrameNet: Theory and practice.

M. Kouylekov and B. Magnini. 2005. Recognizing textual entailment with tree edit distance algorithms. In *Proceedings of the First PASCAL Challenges Workshop on Recognizing Textual Entailment*, 17–20. Southampton, UK.

M. T. Pazienza, M. Pennacchiotti, and F. M. Zanzotto. 2005. Textual entailment as syntactic graph distance: A rule based and a SVM based approach. In *Proceedings of the First PASCAL Challenges Workshop on Recognizing Textual Entailment*, 25–28. Southampton, UK.

Marta Tatu and Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*, 371–378. Vancouver, British Columbia, Canada.

Mehmet Ali Yatbaz. 2008. RTE4: Normalized dependency tree alignment using unsupervised n-gram word similarity score. In *Proceedings of the Fourth PASCAL Challenges Workshop on Recognizing Textual Entailment*. Gaithersburg, Maryland, USA.

Peter Clark and Phil Harrison. 2008. Recognizing textual entailment with logic inference. In *Proceedings of the Fourth PASCAL Challenges Workshop on Recognizing Textual Entailment*. Gaithersburg, Maryland, USA.

R. de Salvo Braz, R. Girju, V. Punyakanok, D. Roth, and M. Sammons. 2005. Textual entailment recognition based on dependency analysis

Table 5: The ablation test results of the UB.dmirg 2010 runs. The uparrow (↑) shows that the result of the ablation test is higher than that of the main run, the downarrow (↓) shows the opposite effect, and the line (|) shows no change in any direction.

Run	Micro-averaged			Macro-averaged per topic		
	precision	recall	f-measure	precision	recall	f-measure
UB.dmirg1_abl-1	12.72↑	9.84↓	11.10↓	13.57↑	10.76↓	12.00↓
UB.dmirg1_abl-2	11.36↓	14.92↑	12.90↑	11.45↓	15.33↑	13.10↑
UB.dmirg2_abl-1	16.00↓	8.89|	11.43↓	15.50↓	9.54↑	11.81↓
UB.dmirg2_abl-2	16.50↓	11.96↑	13.87↑	17.82↓	12.33↑	14.58↑
UB.dmirg3_abl-1	21.76↑	18.10↓	19.76↑	32.80↑	18.88↓	23.96↑
UB.dmirg3_abl-2	13.32↑	47.30↓	20.79↑	14.41↑	48.52↓	22.22↑

and WordNet. In *Proceedings of the First PASCAL Challenges Workshop on Recognizing Textual Entailment*, 29–32. Southampton, UK.

W. E. Bosma and C. Callison-Burch. 2006. Paraphrase substitution for recognizing textual entailment. In *Working Notes of CLEF 2006*, 1–8. Alicante, Spain.

Bahadorreza Ofoghi and John Yearwood. 2010. Learning parse-free event-based features for textual entailment recognition. In *Proceedings of the 23rd Australasian Joint Conference on Artificial Intelligence*, 184–193. Adelaide, Australia.