A Joint Syntactic-Semantic Representation for Recognizing Textual Relatedness

Rui Wang
Saarland University
Saarbruecken, Germany
rwang@coli.uni-sb.de

Yi Zhang, Guenter Neumann
DFKI GmbH
Saarbruecken, Germany
yzhang@coli.uni-sb.de, neumann@dfki.de

Abstract

This paper describes our participation in the *Recognizing Textual Entailment* challenge (RTE-5) in the *Text Analysis Conference* (TAC 2009). Following the two-stage binary classification strategy, our focus this year is to recognize *related* Text-Hypothesis pairs instead of *entailment* pairs. In particular, we propose a joint syntactic-semantic representation to better capture the key information shared by the pair, and also apply a co-reference resolver to group cross-sentential mentionings of the same entities together. For the evaluation, we achieve 63.7% of accuracy on the three-way test, 68.5% on the entailment vs. non-entailment test, and 74.3% on the relatedness recognition. Based on the error analysis, we will work on differentiating entailment and contradiction in the future.

1 Introduction

Recent research on *recognizing textual entailment* (RTE – Dagan et al., 2006; Giampiccolo et al., 2007; Giampiccolo et al., 2008) extends the two-way annotation into three-way[1], making the task more difficult, but more linguistic motivated. The straightforward strategy is to tackle the three-way task is to treat it as a three-way classification task, such that a classifier directly tries to assign one of the following results, *entailment*, *contradiction*, or *unknown*. However, it seems that the performance suffers a significant drop even when using the same classifier and feature model. As mentioned in Wang and Zhang (2009), common approaches based on overlapping information between *text* (T) and *hypothesis* (H) usually over-cover the *entailment* (E) cases, which include the *contradiction* (C) cases as well. They suggest a *Textual Relatedness* measurement and also perform a detailed comparison of different classification strategies, which indicates that identifying *related* (R) and *unknown* (U) cases first might be the most appropriate choice. Following this line of research, we show the relationship between these different cases in Figure 1. In this work, we continue the two-stage classification strategy, that is to make a decision between R and U first and followed by a second decision between

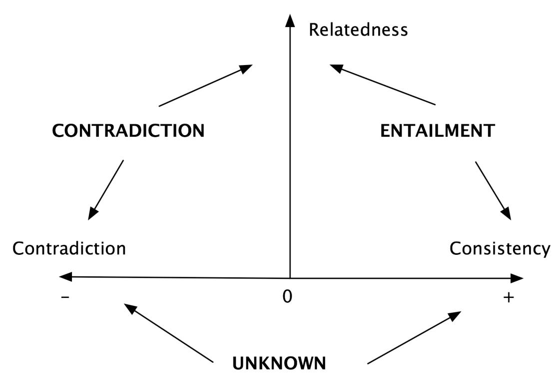

Figure 1 The Relationship between the Three Textual Relations
We treat relatedness as another dimension besides entailment and contradiction. Given a related T-H pair, we further decide whether there is an entailment relation or contradiction in between; if it is unrelated, it will be classified as unknown. Note that in fact these relations cannot cover the whole area (e.g. the directionality of the entailment relation is ignored here), this is just a simplified figure roughly showing the two dimensions.

[1] http://nlp.stanford.edu/RTE3-pilot/ and http://www.nist.gov/tac/tracks/2008/rte/rte.08.guidelines.html

E and C, instead of a direct classification among E, C, and U in one step. And we focus on the first stage.

Apart from the choice of different strategies, there is another issue that has not been addressed in detail in (Wang and Zhang, 2009), that is, the selection of the meaning representation for the text. In fact, any representation is limited in someway, the Bag-of-Words (BoW) representation is an extremely naïve model in this case, and it is insufficient to capture the meaning of text in general. Our previous work (Wang and Neumann, 2007) proposed a pure syntactic approach by only performing a syntactic dependency analysis on corresponding T-H pairs. It utilizes subsequence kernel to capture different variations of the syntactic transformation based on Part-of-Speech (POS) and syntactic dependency path. Although this approach only uses syntactic information, it achieved promising performance. Of course, it is still not a systematic way of handling entailment recognition, which is usually at or beyond the semantic level. Wang and Zhang (2009) work on the semantic dependency graphs, which can deal with (some) syntactic variations like active/passive voice transformation, nominalization of the events, etc. However, the semantic dependency fails to reach the syntactic object inside each prepositional phrase, which is of great importance for matching key information between T and H. Moreover, sentence-based syntactic and semantic dependency analysis will suffer from unsolved cross-sentential co-references, when T contains more than one sentence, and the problem becomes even more severe if the length of T increases (as in RTE-5).

In this work, in order to solve the problems mentioned above,

- We combine syntactic and semantic dependency structure into a connected graph, achieving a new joint representation which can better capture the overlapping information between T and H;
- We also use a co-reference resolver to group different mentionings of the same entity together to share the information between sentences.

The experiment results show that our approach is effective on relatedness recognition, therefore, favors three-way classification more than the traditional two-way annotation, entailment vs. non-entailment. The joint representation, the co-reference resolver, and the richer feature model for the backup strategies all contribute to the final results, while the contribution from different lexical resources is less significant.

In the following, we will firstly mention some related work; Section 3 will introduce our joint representation; Section 4 will describe the matching/alignment module of our system; The experiment results as well as ablation tests of excluding different lexical semantics resources will be presented in Section 5; and finally in Section 6, we will conclude the paper and point out future work.

2 Related Work

Although the term of *Textual Relatedness* has not been widely used by the community, many researchers have already incorporated modules to tackle it, which are usually implemented as an alignment module before the inference/learning module is applied. For example, Pado et al. (2009) mentioned two alignment modules: one is a phrase-based alignment system called MANLI (MacCartney et al., 2008), and the other is a stochastic aligner based on dependency graphs.

As for the whole RTE task, many people directly do the three-way classification with selective features (e.g. Agichtein et al., 2009) or different inference rules to identify entailment and contradiction simultaneously (e.g. Clark and Harrison, 2009); while other researchers also extend their two-way classification system into three-way by performing a second-stage classification afterwards. An interesting task proposed by de Marneffe et al. (2008) suggested an alternative way to deal with the three-way classification, that is, to identify the contradiction cases first. However, it has been shown to be more difficult than the entailment recognition.

Wang and Zhang (2009) performed a detailed comparison of different two-stage binary classification strategies. They claimed that approaches based on overlapping information between T and H actually capture the related text pairs instead of entailment pairs, which is confirmed by experiments. In this work, we will follow the strategy of splitting related/unknown cases out first, but we propose a joint syntactic-

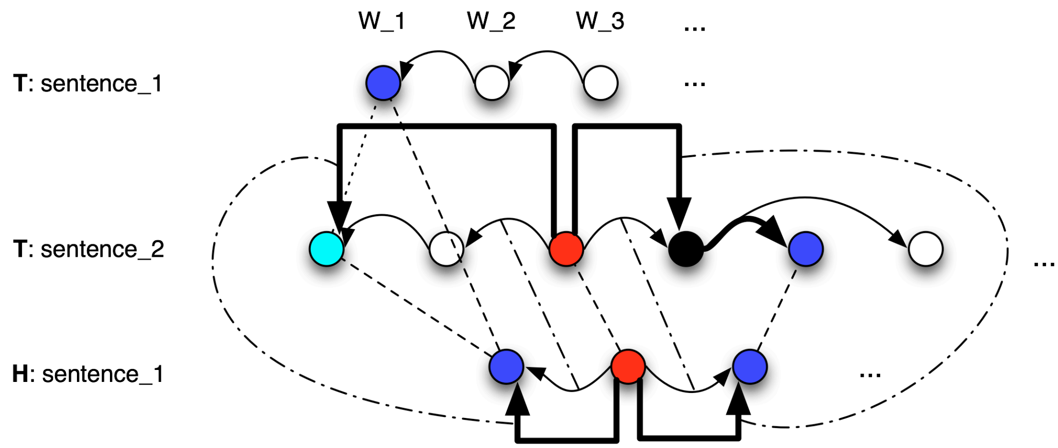

Figure 2 Example of an Alignment

Each circle represents a word/token in the sentence; circles with the same color are aligned word pairs; the light blue circle represents a co-reference of the dark blue circle in sentence 1 of T; each arrow represents a dependency between two words, either a syntactic dependency (curved) or a semantic dependency (orthogonal); a dashed line means an alignment between two words, and a dash-doted line means an alignment between two dependencies. Note that the black circle is how "far" we can get by semantic dependency, which cannot reach the aimed dark blue circle.

semantic representation for text, as opposite to the pure semantic dependency graph as in (Wang and Zhang, 2009).

Basically our previous year's system focusing on NE-based entailment (Wang and Neumann, 2009) could potentially be used for relatedness recognition as well. However, due to the time constraint, we did not integrate the two systems into a unified framework. After all, the old system focuses on two-way entailment recognition (directional rules were applied), and the new system handles relatedness better, which is based on overlapping information shared between T and H.

3 The Joint Representation

Before introducing the joint representation, let us first take a closer look at the problems of the pure syntactic or semantic dependency structure as the meaning representation. Figure 2 shows an abstract example of an alignment between T and H.

We simplified H into a concise sentence with only three words, e.g. a predicate (in red) with a subject and an object, while T contains two sentences (hence T_1 and T_2) and more information (ir)relevant to H. We assume that T_2 is aligned with H with more overlapping words (denoted by the circles with same blue color[2]).

Besides the word alignment, we also check the overlapping syntactic dependency triples (i.e. <word, relation, word>), but we observe that these overlapping syntactic dependency triples cannot help us to reach the aligned words (on the syntactic dependency tree). Therefore, we need to go one level deeper to the semantic dependencies.

Although we could get the left-hand side aligned fully by semantic dependencies, the blue circle in T_2 on right-hand side still cannot be reached. The black circle here is the end of the semantic dependency graph, and usually it is realized as a preposition. Consequently, we take the syntactic dependency into account, which links the black circle to the blue circle and it can be used as a "backup" link for the semantic dependencies.

Therefore, the joint representation consists of two parts: 1) the semantic dependencies (which can be a bag of isolated graphs in some cases); and 2) the syntactic dependencies connecting the content words, where the semantic graph ends at functional words. This is marked in bold in Figure 2.

[2] The light blue circle will be discussed later. For now, we just treat it the same as the dark blue ones.

The light blue circle denotes a co-reference of the dark blue circle in T_1. This occurs more frequently, if the text of T becomes longer. Since we can easily find the alignment between the first word of H and the first word in T_1, the alignment can be potentially passed to the first word in T_2 (which is, for example, a pronoun). Therefore, we apply a co-reference resolution toolkit, BART (Versley et al., 2008) to gather such cross-sentential references. Briefly, the resolver will assign a label for a bag of different mentionings of the same entity it discovers in the text and we just group all the mentionings together (according to the labels) for the word alignment module.

Based on these two processes, we can thus integrate all the information (under a certain discourse) into one unified representation, both horizontally (from different sentences) and vertically (from different levels of linguistic analyses).

The advantage of such representation will be shown in the following example,

T: At least five people have been killed in a head-on train collision **in north-eastern France,** while others are still trapped in the wreckage. All the victims are *adults*.
H: A **French** train crash killed *children*.

This is an example of a contradiction, where the only contradictive part lies on "adults" in T and "children" in H (shown in italics). As being mentioned in Wang and Zhang (2009), this pair can be solved by matching the semantic dependency graphs. However, notice that, "in north-eastern France" in T and "French" in H (shown in bold) cannot be aligned by only syntactic or semantic dependencies, because the AM-LOC argument of the predicate "collision" is the preposition "in", and "France" cannot be reached on the semantic dependency graph. The link from the preposition "in" to the object "France" is a syntactic dependency, which will be included by the joint representation. Similarly, the active/passive voice transformation can also be captured. For example,

T: Yigal Amir, the student who **assassinated** Israeli Prime Minister Yitzhak Rabin, …
H: Yitzhak Rabin was killed **by Yigal Amir.**

"Yigal Amir" in T will be linked to "assassinated" via semantic dependency (the syntactic dependency is not direct); while "Yigal Amir" in H is under the preposition "by" on the syntactic tree, but hidden on the semantic dependency graph.

4 The Matching Module

Basically, we use the same matching algorithm of Wang and Zhang (2009). As we mentioned in the introduction, we break down the three-way classification into a two-stage binary classification and focus on the first stage as a subtask of the main task, which is to determine whether H is related to T. Similar to the probabilistic entailment score, we use a relatedness score to measure such relationship. Due to the nature of the entailment recognition that H should be fully entailed by T, we also make this relatedness relationship similar. Roughly speaking, this relatedness function R(T, H) can be described as whether or how relevant H is to some part of T. In practice, the relevance can be realized as surface string similarity, semantic similarity, or co-occurrence-based similarity.

The only difference here is that, the meaning representation for the new system is the joint syntactic-semantic dependency graph introduced in the previous section instead of the pure semantic dependency graph. Therefore, when we find the best matching sentence pair from T and H, we will (in most of the cases) have two connected graphs at hand. And then, the same decomposition algorithm will be performed to divide the graphs into dependency triples (i.e. <predicate, relation, argument> for semantic dependencies and <parent, relation, child> for syntactic dependencies).

To align the words, we need lexical semantic resources. Fortunately, many people have done research on semantic relatedness in lexical semantic studies. Therefore, these functions can be realized by different string matching algorithms and/or lexical resources. Since the meaning of relevance is rather broad, apart from the string matching of the lemmas, we also incorporate various resources, from distributionally collected ones to handcrafted ontologies. We choose VerbOcean (Chklovski and Pantel, 2004) to obtain the relatedness between predicates (after using WordNet – (Miller, 1993) to change all the nominal predicates into verbs) and use WordNet for the argument alignment. For the verb relations in VerbOcean, we consider all of them as related; and for WordNet, we not only use the synonyms, hyponyms, and hypernyms, but antonyms as well. In addition, the *Normalized Google Distance* (NGD – Cilibrasi and Vitanyi, 2007) is applied to both cases[3] and we use empirical value 0.5 as the threshold.

In all, the main idea here is to incorporate both distributional semantics and ontological semantics in order to see whether their contributions are overlapping or complementary. In order to achieve a better coverage, we use the OR operator to connect all the lexical relatedness functions, which means, if any of them holds, the two items are related.

Finally, besides giving one single answer (whether T and H is related or not), the module provides more outputs, like the alignment ratio of the predicates over all the triples, whether all the predicates in H are aligned, and the same for the arguments. Furthermore, these outputs are used as part of the feature model of the backup strategy, which essentially incorporate features from all linguistic processing levels, BoW, syntactic dependencies, semantic dependencies, and the joint representation.

5 Results and Discussion

We have submitted three runs for the three-way RTE challenge, which have different configurations as follows,

- Run1: Wang and Zhang's system + a backup strategy using BoW, and syntactic dependency features
- Run2: the main system (lenient[4]) + a backup strategy using features from the BoW, syntactic dependency, and semantic dependency
- Run3: the main system (strict) + a backup strategy using features from the BoW, syntactic dependencies, semantic dependencies, and the joint representation

For syntactic dependency parsing, we use the open source MSTParser (McDonald et al., 2005), trained on the Wall Street Journal Sections of the Penn Treebank, using a projective decoder with second-order features. And for the semantic dependency parsing, we use the semantic role labeler described in (Zhang et al., 2008). The system is also trained on the Wall Street Journal sections of the Penn Treebank using PropBank and NomBank annotation of verbal and nominal predicates, and relations to their arguments (Mihai et al., 2008).

The three-way results and the ablation test results are shown in the following Table 1[5],

Runs	Main	Main -VO	Main -WN	Main -VO-WN
DFKI1	50.7%	50.5%	50.7%	50.5%
DFKI2	**63.7%**	63.2%	63.3%	63.0%
DFKI3	63.5%	63.3%	63.3%	63.3%

Table 1 Official Submission Results (three-way)

Compared to Wang and Zhang (2009) system (DFKI1), the improvement of system DFKI2 and DFKI3 is obvious. We attribute it to two reasons: 1) the backup strategies using richer feature models (i.e. including semantic dependency features) contribute to the final results; and furthermore 2) the joint representation of syntactic and semantic dependency and co-reference resolution are also effective. We will take a closer look at these two aspects in the following Table 2, which shows the T-H pairs on which the main approach has a higher confidence,

[3] You may find the NGD values of all the content word pairs in RTE-3, RTE-4, and RTE-5 datasets at http://www.coli.uni-sb.de/~rwang/resources/RTE3_RTE4_NGD.zip and http://www.coli.uni-sb.de/~rwang/resources/RTE5_NGD.zip

[4] We use the same settings of the Wang and Zhang's system. Here, "lenient" is "yyn" according to their definition, meaning either predicate trees ask for a full match or argument trees ask for a full match; and similarly, "strict" is "nny".

[5] VO stands for VerbOcean; WN stands for WordNet; and "-" here means taking the resource(s) out.

Runs	Covered		
	# of T-H Pairs	Main	Backup
DFKI1	102	55.9%	53.9%
DFKI2	102	67.6%	67.6%
DFKI3	10	80.0%	80.0%

Table 2 Results on the Covered Data

We observe that, in fact, the backup strategies determines the final results, since the coverage of the main approach occupies only 1/6 of the dataset, especially for DFKI3, the strict setting of our system only covers 10 T-H pairs. The big gap between DFKI1 and the other two suggests that including richer semantic dependency features does help a lot.

We also calculate the confusion matrix for the three-way submission DFKI2 as follows,

DFKI2		Gold-Standard			
		E	C	U	Total
System	E	238	**60**	77	375
	C	4	21	10	35
	U	58	9	123	190
	Total	300	90	210	600

Table 3 Confusion Matrix of DFKI2 Submission

Although the system confuses between many entailment and unknown cases, the most serious problem seems to be the contradiction recognition, whose recall is the lowest (21/90=23.3%). In fact, this difficulty has been mentioned in the previous research (de Marneffe et al., 2008). Therefore, the differentiation between entailment and contradiction in the second stage classification will be our future focus.

Finally, we also present our two-way results, both on the traditional two-way classes (Table 4) and related vs. unknown classes (Table 5).

Runs	Main	Main -VO	Main -WN	Main -VO-WN
DFKI1	62.5%	62.5%	62.7%	62.5%
DFKI2	66.8%	66.5%	66.7%	66.3%
DFKI3	**68.5%**	68.3%	68.3%	68.3%

Table 4 Results of Entailment vs. Non-Entailment

Runs	Main	Main -VO	Main -WN	Main -VO-WN
DFKI1	74.0%	73.7%	73.8%	73.7%
DFKI2	**74.3%**	73.7%	73.8%	73.5%
DFKI3	72.3%	72.2%	72.2%	72.2%

Table 5 Results of Related vs. Unknown

These two tables clearly show that all our three runs do well for the relatedness recognition, which meets our original goal. The overall improvement from the worst to the best results on the traditional two-way annotation is less than the three-way annotation (6% vs. 13%).

Notice that the setting with the best three-way result (DFKI2) is different from the setting of the best two-way result (DFKI3). It seems that, a model of richer features would contribute more to the two-way task; while a simplified version would give us a better three-way result.

6 Conclusion and Future Work

To sum up, the results of this year's participation is encouraging. Given that the task is more difficult (longer Ts), we still achieve better results than last year's results (although on different data sets). Especially, our approach based on the joint syntactic-semantic structure is effective for the three-way task, which we view as a more linguistically motivated annotation scheme. This also confirms the claim that methods based on overlapping information between T and H does a better job on relatedness recognition comparing with the entailment recognition.

For the future work, as we mentioned in the previous section, we plan to do more fine-grained analysis on the second-stage classification, that is, entailment vs. contradiction. And the directionality of entailment should also be taken into consideration.

Acknowledgments

The first author thanks the PIRE PhD scholarship program; the second author thanks DFKI and the Cluster of Excellence on Multimodal Computing and Interaction for their support of the work; and this work presented here was partially supported by the EC-funded project QALL-ME.

References

Agichtein, E., Askew, W., and Liu, Y. 2009. Combining Lexical, Syntactic, and Semantic Evidence for Textual Entailment Classification. In *Proceedings of the First Text Analysis Conference (TAC 2008)*.

Chklovski, T. and Pantel, P. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04),* volume 34, pages 343-360, Barcelona, Spain.

Cilibrasi, R. and Vitanyi, P. 2007. The Google Similarity Distance. In *IEEE/ACM Transactions on Knowledge and Data Engineering*, 19(3):370–383.

Clark, P. and Harrison, P. 2009. Recognizing Textual Entailment with Logical Inference. In *Proceedings of the First Text Analysis Conference (TAC 2008)*.

Dagan I., Glickman, O., and Magnini, B. 2006. The PASCAL Recognising Textual Entail- ment Challenge. In *Machine Learning Challenges, volume 3944 of Lecture Notes in Computer Science*, pages 177–190. Springer.

de Marneffe, M., Rafferty A., and Manning, C. 2008. Finding contradictions in text. In *Proceedings of ACL-HLT 2008*.

Giampiccolo, D., Dang, H., Magnini, B., Dagan, I., Cabrio, E., and Dolan, B. 2009. The Fourth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the First Text Analysis Conference (TAC 2008)*.

Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B. 2007. The Third PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the ACL Workshop on Textual Entailment and Paraphrasing*, Prague.

McDonald, R., Pereira, F., Ribarov, K., Hajic, J. 2005. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of HLT-EMNLP 2005*, Vancouver, Canada, pp. 523–530.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. 1993. Five papers on WordNet. *Technical report*, Cognitive Science Laboratory, Princeton University.

Pado, S., de Marneffe, M., MacCartney, B., Rafferty, A., Yeh, E., and Manning, C. 2009. Deciding entailment and contradiction with stochastic and edit distance-based alignment. In *Proceedings of the First Text Analysis Conference (TAC 2008)*.

Surdeanu, M., Johansson, R., Meyers, A., Ma`rquez, L., and Nivre, J. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL- 2008)*, Manchester, UK.

Versley, Y., Ponzetto, S., Poesio M., Eidelman, V., Jern, A., Smith, J., Yang, X., and Moschitti, A. 2008. BART: A Modular Toolkit for Coreference Resolution. In *ACL2008 Demonstration session*.

Wang, R. and Neumann, G. 2007. Recognizing Textual Entailment Using a Subsequence Kernel Method. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence (AAAI-07)*, July 22-26, Vancouver, Canada.

Wang, R. and Neumann, G. 2009. An Accuracy-Oriented Divide-and-Conquer Strategy for Recognizing Textual Entailment. In *Proceedings of the Text Analysis Conference (TAC 2008) Workshop - RTE-4 Track*, Gaithersburg, Maryland, USA, NIST.

Wang, R. and Zhang, Y. 2009. Recognizing Textual Relatedness with Predicate-Argument Structures. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2009),* Singapore, Singapore.

Zhang Y., Wang R., and Uszkoreit, H. 2008. Hybrid Learning of Dependency Structures from Heterogeneous Linguistic Resources. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL 2008)*, Pages 198-202, Manchester, United Kingdom, ACL.