Recognizing Implied Predicate-Argument Relationships in Textual Inference

Asher Stern
Computer Science Department
Bar-Ilan University
astern7@cs.biu.ac.il

Ido Dagan
Computer Science Department
Bar-Ilan University
dagan@cs.biu.ac.il

Abstract

We investigate recognizing implied predicate-argument relationships which are not explicitly expressed in syntactic structure. While prior works addressed such relationships as an extension to semantic role labeling, our work investigates them in the context of textual inference scenarios. Such scenarios provide prior information, which substantially eases the task. We provide a large and freely available evaluation dataset for our task setting, and propose methods to cope with it, while obtaining promising results in empirical evaluations.

1 Motivation and Task

This paper addresses a typical sub-task in textual inference scenarios, of recognizing implied predicate-argument relationships which are not expressed explicitly through syntactic structure. Consider the following example:

(i) The crucial role Vioxx plays in Merck's portfolio was apparent last week when Merck's shares plunged 27 percent to 33 dollars after the withdrawal announcement.

While a human reader understands that the withdrawal refers to Vioxx, and hence an implied predicate-argument relationship holds between them, this relationship is not expressed in the syntactic structure, and will be missed by syntactic parsers or standard semantic role labelers.

This paper targets such types of implied relationships in textual inference scenarios. Particularly, we investigate the setting of *Recognizing Textual Entailment (RTE)* as a typical scenario of textual inference. We suggest, however, that the same challenge, as well as the solutions proposed in our work, are applicable, with proper adaptations, to other textual-inference scenarios, like *Question Answering*, and *Information Extraction* (see Section 6).

An RTE problem instance is composed of two text fragments, termed *Text* and *Hypothesis*, as input. The task is to recognize whether a human reading the Text would infer that the Hypothesis is most likely true (Dagan et al., 2006). For our problem, consider a positive Text Hypothesis pair, where the Text is example (i) above and the Hypothesis is:

(ii) Merck withdrew Vioxx.

A common approach for recognizing textual entailment is to verify that all the textual elements of the Hypothesis are *covered*, or *aligned*, by elements of the Text. These elements typically include lexical terms as well as relationships between them. In our example, the Hypothesis lexical terms ("Merck", "withdrew" and "Vioxx") are indeed covered by the Text. Yet, the predicate-argument relationships (e.g., "withdrawal-Vioxx") are not expressed in the text explicitly. In such a case, an RTE system has to verify that the predicate-argument relationships which are explicitly expressed in the Hypothesis, are *implied* from the Text discourse. Such cases are quite frequent ($\sim$17%) in the settings of our dataset, described in Section 3.

Consequently, we define the task of recognizing implied predicate-argument relationships, with illustrating examples in Table 1, as follows. The input includes a *Text* and a *Hypothesis*. Two terms in the Hypothesis, *predicate* and *argument*, are marked, where a predicate-argument relationship between them is explicit in the Hypothesis syntactic structure. Two terms in the Text, *candidate-predicate* and *candidate-argument*, aligned to the Hypothesis predicate and argument, are marked as well. However, no predicate-argument relationship between them is expressed syntactically. The task is to recognize whether the predicate-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 739–744,
Baltimore, Maryland, USA, June 23-25 2014.

#	*Hypothesis*	*Text*	*Y/N*
1	Merck [withdrew]$_{\text{pred}}$ [Vioxx]$_{\text{arg}}$ from the market.	The crucial role [Vioxx]$_{\text{cand-arg}}$ plays in Merck's portfolio was apparent last week when Merck's shares plunged 27 percent to 33 dollars after the [withdrawal]$_{\text{cand-pred}}$ announcement.	Y
2	Barbara Cummings heard the tale of a woman who was coming to Crawford to [join]$_{\text{pred}}$ Cindy Sheehans [protest]$_{\text{arg}}$.	Sheehan's [protest]$_{\text{cand-arg}}$ is misguided and is hurting troop morale. ... Sheehan never wanted Casey to [join]$_{\text{cand-pred}}$ the military.	N
3	Casey Sheehan was [killed]$_{\text{pred}}$ in [Iraq]$_{\text{arg}}$.	5 days after he arrived in [Iraq]$_{\text{cand-arg}}$ last year, Casey Sheehan was [killed]$_{\text{cand-pred}}$.	Y
4	Hurricane Rita [threatened]$_{\text{pred}}$ [New Orleans]$_{\text{arg}}$.	Hurricane Rita was upgraded from a tropical storm as it [threatened]$_{\text{cand-pred}}$ the southeastern United States, forcing an alert in southern Florida and scuttling plans to repopulate [New Orleans]$_{\text{cand-arg}}$ after Hurricane Katrina turned it into a ghost city 3 weeks earlier.	Y
5	Alberto Gonzales defends [renewal]$_{\text{pred}}$ of the [Patriot Act]$_{\text{arg}}$ to Congress.	A senior official defended the [Patriot Act]$_{\text{cand-arg}}$ ... ... President Bush has urged Congress to [renew]$_{\text{cand-pred}}$ the law ...	Y
6	The [train]$_{\text{arg}}$ [crash]$_{\text{pred}}$ injured nearly 200 people.	At least 10 people were killed ... in the [crash]$_{\text{cand-pred}}$ ... Alvarez is accused of ... causing the derailment of one [train]$_{\text{cand-arg}}$ ...	Y

Table 1: Example task instances from our dataset. The last column specifies the Yes/No annotation, indicating whether the sought predicate-argument relationship is implied in the Text. For illustration, a dashed line indicates an explicit argument that is related to the candidate argument through some kind of discourse reference. Pred, arg and cand abbreviate predicate, argument and candidate respectively.

argument relationship, as expressed in the Hypothesis, holds implicitly also in the Text.

To address this task, we provide a large and freely available annotated dataset, and propose methods for coping with it. A related task, described in the next section, deals with such implied predicate-argument relationships as an extension to *Semantic Role Labeling*. While the results reported so far on that annotation task were relatively low, we suggest that the task itself may be more complicated than what is actually required in textual inference scenarios. On the other hand, the results obtained for our task, which does fit textual inference scenarios, are promising, and encourage utilizing algorithms for this task in actual inference systems.

2 Prior Work

The most notable work targeting implied predicate-argument relationships is the 2010 SemEval task of *Linking Events and Their Participants in Discourse* (Ruppenhofer et al., 2009). This task extends *Semantic Role Labeling* to cases in which a core argument of a predicate is missing in the syntactic structure but a filler for the corresponding semantic role appears elsewhere and can be inferred from discourse. For example, in the following sentence the semantic role *goal* is unfilled:

(iii) He arrived (0^{Goal}) at 8pm.

Yet, we can expect to find an implied filler for *goal* elsewhere in the document.

The SemEval task, termed henceforth as *Implied SRL*, involves three major sub-tasks. First, for each predicate, the unfilled roles, termed *Null Instantiations (NI)*, should be detected. Second, each NI should be classified as *Definite NI (DNI)*, meaning that the role filler must exist in the discourse, or *Indefinite NI* otherwise. Third, the DNI fillers should be found (DNI linking).

Later works that followed the SemEval challenge include (Silberer and Frank, 2012) and (Roth and Frank, 2013), which proposed auto-

matic dataset generation methods and features which capture discourse phenomena. Their highest result was 12% F1-score. Another work is the probabilistic model of Laparra and Rigau (2012), which is trained by properties captured not only from implicit arguments but also from explicit ones, resulting in 19% F1-score. Another notable work is (Gerber and Chai, 2012), which was limited to ten carefully selected nominal predicates.

2.1 Annotations vs. Recognition

Comparing to the implied SRL task, our task may better fit the needs of textual inference. First, some relatively complex steps of the implied SRL task are avoided in our setting, while on the other hand it covers more relevant cases.

More concretely, in textual inference the candidate predicate and argument are typically identified, as they are aligned by the RTE system to a predicate and an argument of the Hypothesis. Thus, the only remaining challenge is to verify that the sought relationship is implied in the text. Therefore, the sub-tasks of identifying and classifying DNIs can be avoided.

On the other hand, in some cases the candidate argument is not a DNI, but is still required in textual inference. One type of such cases are non-core arguments, which cannot be *Definite NIs*. However, textual inference deals with non-core arguments as well (see example 3 in Table 1).

Another case is when an implied predicate-argument relationship holds even though the corresponding role is already filled by another argument, hence not an NI. Consider example 4 of Table 1. While the object of "threatened" is filled (in the Text) by "southeastern United States", a human reader also infers the "threatened-New Orleans" relationship. Such cases might follow a meronymy relation between the filler ("southeastern United States") and the candidate argument ("New Orleans"), or certain types of discourse (co-)references (e.g., example 5 in Table 1), or some other linguistic phenomena. Either way, they are crucial for textual inference, while not being NIs.

3 Dataset

This section describes a semi-automatic method for extracting candidate instances of implied predicate-argument relationship from an RTE dataset. This extraction process directly follows our task formalization. Given a Text Hypothesis pair, we locate a predicate-argument relationship in the Hypothesis, where both the predicate and the argument appear also in the Text, while the relationship between them is not expressed in its syntactic structure. This process is performed automatically, based on syntactic parsing (see below). Then, a human reader annotates each instance as "Yes" – meaning that the implied relationship indeed holds in the Text, or "No" otherwise. Example instances, constructed by this process, are shown in Table 1.

In this work we used lemma-level lexical matching, as well as nominalization matching, to align the Text predicates and arguments to the Hypothesis. We note that more advanced matching, e.g., by utilizing knowledge resources (like WordNet), can be performed as well. To identify *explicit* predicate-argument relationships we utilized dependency parsing by the Easy-First parser (Goldberg and Elhadad, 2010). Nominalization matching (e.g., example 1 of Table 1) was performed with Nomlex (Macleod et al., 1998).

By applying this method on the RTE-6 dataset (Bentivogli et al., 2010), we constructed a dataset of 4022 instances, where 2271 (56%) are annotated as positive instances, and 1751 as negative ones. This dataset is significantly larger than prior datasets for the implied SRL task. To calculate inter-annotator agreement, the first author also annotated 185 randomly-selected instances. We have reached high agreement score of 0.80 Kappa. The dataset is freely available at `www.cs.biu.ac.il/~nlp/resources/downloads/implied-relationships`.

4 Recognition Algorithm

We defined 15 features, summarized in Table 2, which capture local and discourse phenomena. These features do not depend on manually built resources, and hence are portable to resource-poor languages. Some features were proposed in prior works, and are marked by G&C (Gerber and Chai, 2012) or S&F (Silberer and Frank, 2012). Our best results were obtained with the *Random Forests* learning algorithm (Breiman, 2001). The first two features are described in the next subsection, while the others are explained in the table itself.

4.1 Statistical discourse features

Statistical features in prior works mostly capture general properties of the predicate and the

#	Category	Feature	Prev. work
1	statistical discourse	co-occurring predicate (explained in subsection 4.1)	*New*
2		co-occurring argument (explained in subsection 4.1)	*New*
3	local discourse	co-reference: whether an explicit argument of p co-refers with a.	*New*
4		last known location: If the NE of a is "location", and it is the last location mentioned before p in the document.	*New*
5		argument prominence: The frequency of the lemma of a in a two-sentence windows of p, relative to all entities in that window.	S&F
6		predicate frequency in document: The frequency of p in the document, relative to all predicates appear in the document.	G&C
7	local candidate properties	statistical argument frequency: The Unigram-model likelihood of a in English documents, calculated from a large corpus.	*New*
8		definite NP: Whether a is a definite NP	G&C
9		indefinite NP: Whether a is an indefinite NP	G&C
10		quantified predicate: Whether p is quantified (i.e., by expressions like "every ...", "a good deal of ...", etc.)	G&C
11		NE mismatch: Whether a is a named entity but the corresponding argument in the hypothesis is not, or vice versa.	*New*
12	predicate-argument relatedness	predicate-argument frequency: The likelihood of a to be an argument of p (formally: $Pr(a\|p)$) in a large corpus.	similar feature in G&C
13		sentence distance: The distance between p and a in sentences.	G&C, S&F
14		mention distance: The distance between p and a in entity-mentions.	S&F
15		shared head-predicate: Whether p and a are themselves arguments of another predicate.	G&C

Table 2: Algorithmic features. p and a denote the candidate predicate and argument respectively.

argument, like selectional preferences, lexical similarities, etc. On the contrary, our statistical features follow the intuition that *explicit* predicate-argument relationships in the discourse provide plausible indication that an *implied* relationship holds as well. In our experiments we collected the statistics from Reuters corpus RCV1 (`trec.nist.gov/data/reuters/reuters.html`), which contains more than 806,000 documents.

We defined two features: *Co-occurring predicate* and *Co-occurring argument*. Let p and a be the candidate predicate and the argument in the text. While they are not connected syntactically, each of them often has an explicit relationships with *other* terms in the text, that might support the sought (implied) relationship between a and p.

More concretely, a is often an *explicit* argument of another predicate p'. For example, example 6 in Table 1 includes the explicit relationship "derailment of train", which might indicate the implied relationship "crash of train". Hence p="crash", a="train" and p'="derailment". The *Co-occurring predicate* feature estimates the probability that a document would contain a as an argument of p, given that a appears elsewhere in that document as an argument of p', based on explicit predicate-argument relationships in a large corpus.

Similarly, the *Co-occurring argument* feature captures cases where p has another *explicit* argument, a'. This is exemplified in example 5 of Table 1, where p="renew", a="Patriot Act" and a'="law". Accordingly, the feature quantifies the probability that a document including the relationship p-a' would also include the relationship p-a.

More details about these features can be found in the first author's Ph.D. thesis at `www.cs.biu.ac.il/~nlp/publications/theses/`

5 Results

We tested our method in a cross-validation setting, and obtained high result as shown in the first row of Table 3. Since our task and dataset are novel, there is no direct baseline with which we can compare this result. As a reference point we mention the majority class proportion, and also report a configuration in which only features adopted from prior works (G&C and S&F) are utilized. This

Configuration	Accuracy %	Δ %
Full algorithm	**81.0**	–
Union of prior work	78.0	3.0
Major category (all true)	56.5	24.5
Ablation tests		
no statistical discourse	79.9	1.1
no local discourse	79.3	1.7
no local candidate properties	79.2	1.8
no predicate-argument relatedness	79.7	1.3

Table 3: Accuracy of our method, followed by baselines and ablation tests.

Configuration (input)	Recall	Precision	F1 %
Explicit only	44.6	**44.3**	44.4
Human annotations	**50.9**	43.4	**46.8**
Algorithm recognition	48.5	42.3	45.2

Table 4: RTE-6 Experiment

comparison shows that the contribution of our new features (3%) is meaningful, which is also statistically significant with $p < 0.01$ using *Bootstrap Resampling* test (Koehn, 2004). The high results show that this task is feasible, and its solutions can be adopted as a component in textual inference systems. The positive contribution of each feature category is shown in ablation tests.

An additional experiment tests the contribution of recognizing implied predicate-argument relationships for overall RTE, specifically on the RTE-6 dataset. For the scope of this experiment we developed a simple RTE system, which uses the F1 optimized logistic regression classifier of Jansche (2005) with two features: lexical coverage and predicate-argument relationships coverage. We ran three configurations for the second feature, where in the first only syntactically expressed relationships are used, in the second all the implied relationships, as detected by a human annotator, are added, and in the third only the implied relationships detected by our algorithm are added.

The results, presented in Table 4, first demonstrate the full potential of the implied relationship recognition task to improve textual entailment recognition (Human annotation vs. Explicit only). One third of this potential improvement is achieved by our algorithm[1]. Note that all these results are higher than the median result in the RTE-6 challenge (36.14%). While the delta in the F1 score is small in absolute terms, such magnitudes are typical in RTE for most resources and tools (see (Bentivogli et al., 2010)).

[1]Following the relatively modest size of the RTE dataset, the Algorithm vs. Explicit result is not statistically significant ($p \simeq 0.1$). However, the Human annotation vs. Explicit result is statistically significant with $p < 0.01$.

6 Discussion and Conclusions

We formulated the task of recognizing implied predicate-argument relationships within textual inference scenarios. We compared this task to the labeling task of SemEval 2010, where no prior information about candidate arguments in the text is available. We point out that in textual inference scenarios the candidate predicate and argument are given by the Hypothesis, while the challenge is only to verify that a predicate-argument relationship between these candidates is implied from the given Text. Accordingly, some complex steps necessitated in the SemEval task can be avoided, while additional relevant cases are covered.

Moreover, we have shown that this simpler task is more feasibly solvable, where our 15 features achieved more than 80% accuracy.

While our dataset and algorithm were presented in the context of RTE, the same challenge and methods are applicable to other textual inference tasks as well. Consider, for example, the *Question Answering (QA)* task. Typically QA systems detect a candidate predicate that matches the question's predicate. Similarly, candidate arguments, which match either the expected answer type or other arguments in the question are detected too. Consequently, our methods which exploit the availability of the candidate predicate and argument can be adapted to this scenario as well.

Similarly, a typical approach for *Event Extraction* (a sub task of *Information Extraction*) is to start by applying an entity extractor, which identifies argument candidates. Accordingly, candidate predicate and arguments are detected in this scenario too, while the remaining challenge is to assess the likelihood that a predicate-argument relationship holds between them.

Following this observation, we propose future work of applying our methods to other tasks. An additional direction for future work is to further develop new methods for our task, possibly by incorporating SRL resources and/or linguistically oriented rules, in order to improve the results we achieved so far.

Acknowledgments

This work was partially supported by the EC-funded project EXCITEMENT (FP7ICT-287923).

References

Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2010. The sixth pascal recognizing textual entailment challenge. In *Proccidings of TAC*.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1).

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The stanford typed dependencies representation. In *proceedings of COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*.

Matthew Gerber and Joyce Y. Chai. 2012. Semantic role labeling of implicit arguments for nominal predicates. *Computational Linguistics*.

Yoav Goldberg and Michael Elhadad. 2010. An efficient algorithm for easy-first non-directional dependency parsing. In *Proceedings of NAACL*.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of EMNLP*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explorations*, 11(1).

Martin Jansche. 2005. Maximum expected f-measure training of logistic regression models. In *Proceedings of EMNLP*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*.

Egoitz Laparra and German Rigau. 2012. Exploiting explicit annotations and semantic types for implicit argument resolution. In *Proceedings of IEEE-ICSC*.

Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barrett, and Ruth Reeves. 1998. Nomlex: A lexicon of nominalizations. In *Proceedings of EURALEX*.

Michael Roth and Anette Frank. 2013. Automatically identifying implicit arguments to improve argument linking and coherence modeling. In *Proceedings of *SEM*.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2009. Semeval-2010 task 10: Linking events and their participants in discourse. In *The NAACL-HLT 2009 Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-09)*.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2010. Semeval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*.

Carina Silberer and Anette Frank. 2012. Casting implicit role linking as an anaphora resolution task. In *Proceedings of *SEM*.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL*.