Precision-focused Textual Inference

D. G. Bobrow, C. Condoravdi, R. Crouch, V. de Paiva, L. Karttunen, T. H. King, R. Nairn, L. Price, A. Zaenen
Palo Alto Research Center

Abstract

This paper describes our system as used in the RTE3 task. The system maps premise and hypothesis pairs into an abstract knowledge representation (AKR) and then performs entailment and contradiction detection (ECD) on the resulting AKRs. Two versions of ECD were used in RTE3, one with strict ECD and one with looser ECD.

1 Introduction

In the RTE textual entailment challenge, one is given a source text T and a hypothesis H, and the task is to decide whether H can be inferred from T. Our system interprets inference in a strict way. Given the knowledge of the language embedded in the system, does the hypothesis logically follow from the information embedded in the text? Thus we are emphasizing precision, particularly in question-answering. This was reflected in our results in the RTE3 challenge. We responded correctly with YES to relatively few of the examples, but on the QA-type examples, we achieved 90-95% average precision.

The methodology employed is to use the linguistic information to map T and H onto a logical form in AKR, our Abstract Knowledge Representation. The AKR is designed to capture the propositions the author of a statement is committed to. For the sake of ECD, the representation of T may include elements that are not directly expressed in the text. For example, in the AKR of *John bought a car* includes the fact that the car was sold. The AKR of *John forgot to buy milk* includes the fact that John did not buy milk. Our reasoning algorithm tries to determine whether the AKR of H is subsumed by the AKR of T and detect cases when they are in conflict.

The Entailment and Contradiction Detection (ECD) algorithm makes a distinction that is not part of the basic RTE challenge. If T entails the negation of H, we answer NO (Contradiction). On the other hand, if there is no direct entailment we answer UNKNOWN. We do not try to construct a likely scenario that would link T and H. Nor have we tried to collect data on phrases that would tend to indicate such likely associations between T and H. That approach is clearly very useful (e.g. (Hickl et al., 2006)), and could be used as a backup strategy with our more formal entailment approach. We have chosen to focus on strict structural and lexical entailments.

Process	**Output**
Text-Breaking	Delimited sentences
Named-entity recognition	Type-marked Entities
Morphological Analysis	Word stems plus features
LFG Parsing	Functional Structure
Semantic processing	Scope, Predicate-argument structure
AKR rules	Conceptual, Contextual, Temporal Structure

Figure 1: The processing pipeline: processes with their ambiguity-enabled packed outputs

This paper describes the processing pipeline for mapping to AKR, the ECD algorithm, the challenges we faced in processing the RTE data and a summary of our results on RTE3.

2 Process Pipeline

Figure 1 shows the processing pipeline for mapping texts to AKR. The input is a text of one or more sentences.

All components of the system are "ambiguity enabled" (Maxwell and Kaplan, 1991). This allows each component to accept ambiguous input in a "packed" format, process it without unpacking the ambiguities, and then pass packed input to the next stage. The syntactic component, LFG Parsing, also has a stochastic disambiguation system which allows us to pass the n-best on to the semantics (Riezler et al., 2002); for the RTE3 challenge, we used

Proceedings of the Workshop on Textual Entailment and Paraphrasing, pages 16–21,
Prague, June 2007.

n=50.

The parser takes the output of the morphology (i.e. a series of lemmata with their tags) and produces a tree (constituent-structure) and a dependency structure (functional-structure) represented as an attribute-value matrix. The functional-structure is of primary importance for the semantics and AKR. In particular, it encodes predicate-argument relations, including long-distance dependencies, and provides other syntactic features (e.g. number, tense, noun type).

The output of the syntax is input for the semantics that is produced by an ambiguity enabled packed rewriting system. The semantics is described in detail in (Crouch and King, 2006). Semantic processing assigns scope to scope-bearing elements such as negation and normalizes the output of the syntax. This normalization includes reformulating syntactic passives as actives (e.g. *The cake was eaten by Mary. / Mary ate the cake.*), resolving many null pronouns (e.g. *Laughing, John entered the room / John*$_i$ *laughing, John*$_i$ *entered the room.*), and canonicalizing measure phrases, comparatives, and dates. More complex normalizations involve converting nominal deverbals into the equivalent verbal form, identifying arguments of the verb from the arguments of the nominal (Gurevich et al., 2006). For example, the semantic representation of *Iraq's destruction of its* WMD is similar to the representation of *Iraq destroyed its* WMD.

The final main task of the semantics rules is to convert words into concepts and syntactic grammatical functions into roles. The mapping onto concepts uses WordNet (Fellbaum, 1998) to map words into lists of synsets. The named entity types provided by the morphology and syntax are used to create more accurate mapping of proper nouns since these are not systematically represented in WordNet. The semantic rules use the grammatical function subcategorization information from the verb and the role information found in extended VerbNet (Kipper et al., 2000) to map syntactic subjects, objects, and obliques into more abstract thematic roles such as Agent, Theme, and Goal (Crouch and King, 2005). This mapping into thematic-style roles allows the system to correctly align the arguments in pairs like (1) and (2), something which is impossible using just syntactic functions. In the first, the object and subject have a common thematic role in the alternation between transitive and intransitive; while in the second, the common role is shared by the subjects.

(1) John broke the vase$_{syn:object,sem:patient}$.
The vase$_{syn:subject,sem:patient}$ broke.

(2) John$_{syn:subject,sem:agent}$ ate the cake.
John$_{syn:subject,sem:agent}$ ate.

The goal of these semantic normalizations is to abstract away from the syntactic representation so that sentences with similar meaning have similar semantic representations. However, the semantics is still fundamentally a linguistic level of representation; further abstraction towards the meaning is done in the mapping from semantics to AKR. The AKR is the level of representation that is used to determine entailment and contradiction in our RTE3 system. A preliminary description of its logic was provided in (Bobrow et al., 2005). The AKR mapping converts grammatical tense and temporal modifiers into temporal relations, identifies anaphoric referents and makes explicit the implied relation between complement clauses and the main verb (e.g. for *manage*, *fail*) (Nairn et al., 2006). AKR also deals with standard phrases that are equivalent to simple vocabulary terms. For example, *take a flight to New York* is equivalent to *fly to New York*. These uses of "light" verbs (e.g. *take*, *give*) are not included in synonyms found in WordNet. Another class of phrasal synonyms involve inchoatives (e.g. *take a turn for the worse/worsen*). We included a special set of transformation rules for phrasal synonyms: some of the rules are part of the mapping from semantics to AKR while others are part of the ECD module. The mapping to AKR is done using the same ambiguity-enabled ordered rewriting system that the semantics uses, allowing the AKR mapping system to efficiently process the packed output of the semantics.

The AKR for a sentence like *Bush claimed that Iraq possessed* WMD*s* in Figure 2 introduces two contexts: a top level context t, representing the commitments of the speaker of sentence, and an embedded context claim_cx:37 representing the state of affairs according to Bush's claim. The two contexts are related via the Topic role of the claim event. The representation contains terms like claim:37 or

Conceptual Structure

```
subconcept(claim:37,[claim-1,...,claim-5])
role(Topic,claim:37,claim_cx:37)
role(Agent,claim:37,Bush:1)
subconcept(Bush:1,[person-1])
alias(Bush:1,[Bush])
role(cardinality_restriction,Bush:1,sg)
subconcept(possess:24,[possess-1,own-1,possess-3])
role(Destination,possess:24,wmd:34)
role(Agent,possess:24,Iraq:19)
subconcept(Iraq:19,[location-1,location-4])
alias(Iraq:19,[Iraq])
role(cardinality_restriction,Iraq:19,sg)
subconcept(wmd:34,
   [weapon_of_mass_destruction-1])
role(cardinality_restriction,wmd:34,pl)
```

Contextual Structure

```
context(t)
context(claim_cx:37)
context_relation(t,claim_cx:37,crel(Topic,claim:37))
instantiable(Bush:1,t)
instantiable(Iraq:19,t)
instantiable(claim:37,t)
instantiable(Iraq:19,claim_cx:37)
instantiable(possess:24,claim_cx:37)
instantiable(wmd:34,claim_cx:37)
```

Temporal Structure

```
temporalRel(After,Now,claim:37)
temporalRel(After,claim:37,possess:24)
```

Figure 2: AKR for *Bush claimed that Iraq possessed* WMD*s*.

Bush:1 which refer to the kinds of object that the sentence is talking about. The subconcept facts explicitly link these terms to their concepts in WordNet. Thus claim:37 is stated to be some subkind of the type claim-1, etc., and wmd:34 to be some subkind of the type weapon_of_mass_destruction-1. Terms like claim:37 and wmd:34 do not refer to individuals, but to concepts (or types or kinds). Saying that there is some subconcept of the kind weapon_of_mass_destruction-1, where this subconcept is further restricted to be a kind of WMD possessed by Iraq, does not commit you to saying that there are any instances of this subconcept.

The instantiable assertions capture the commitments about the existence of the kinds of object described. In the top-level context t, there is a commitment to an instance of Bush and of a claim:37 event made by him. However, there is no top-level commitment to any instances of wmd:34 possessed by Iraq:19. These commitments are only made in the embedded claim_cx:37 context. It is left open whether these embedded commitments correspond, or not, to the beliefs of the speaker. Two distinct levels of structure can thus be discerned in AKR: a conceptual structure and a contextual structure. The conceptual structure, through use of subconcept and role assertions, indicates the subject matter. The contextual structure indicates commitments as to the existence of the subject matter via instantiability assertions linking concepts to contexts, and via context relations linking contexts to contexts. In addition, there is a temporal structure that situates the events described with respect to the time of utterance and temporally relates them to one another.

3 Entailment and Contradiction Detection

ECD is implemented as another set of rewrite rules, running on the same packed rewrite system used to generate the AKR representations. The rules (i) align concept and context terms in text (T) and hypothesis (H) AKRs, (ii) calculate concept subsumption orderings between aligned T and H terms, and (iii) check instantiability and uninstantiability claims in the light of subsumption orderings to determine whether T entails H, T contradicts H, or T neither entails not contradicts H. For the purposes of RTE3, both contradiction and neither contradiction nor entailment are collapsed into a NO (does not follow) judgment.

One of the novel features of this approach is that T and H representations do not need to be disambiguated before checking for entailment or contradiction. The approach is able to detect if there is one reading of T that entails (or contradicts) one reading of H. The T and H passages can in effect mutually disambiguate one another through the ECD. For example, although *plane* and *level* both have multiple readings, they can both refer to a horizontal surface, and in that sense *The plane is dry* entails *The level is dry*, and vice versa.

The first phase of ECD aligns concepts and context terms in the T and H AKRs. Concepts are represented as lists of WordNet hypernym lists, in WordNet sense order. Two concept terms can be aligned if a sense synset of one term (i.e. the first element of one of the term's hypernym lists) is contained in a hypernym list of the other term. The alignment can be weighted according to word sense; so a concept overlap on the first senses of a T and H term counts for more than a concept overlap on the *n* and *m*th senses. However, no weightings were used in RTE3. For named entities, alignment demands not only a concept overlap, but also an intersection in the "alias" forms of the proper nouns. For example,"George Bush" may be aligned with "George" or with "Bush". Context alignment relies on associating each context with an indexing concept, usually the concept for the main verb in the clause heading the context. Contexts are then aligned on the basis of these concept indices.

Typically, an H term can align with more than one T term. In such cases all possible alignments are proposed, but the alignment rules put the alternative alignments in different parts of the choice space.

Having aligned T and H terms, rules are applied to determine concept specificity and subsumption relations between aligned terms. Preliminary judgments of specificity are made by looking for hypernym inclusion. For example, an H term denoting the concept "person" is less specific than a T term denoting "woman". These preliminary judgments need to be revised in the light of role restrictions modifying the terms: a "tall person" is neither more nor less specific than a "woman". Revisions to specificity judgments also take into account cardinality modifiers: while "person" is less specific than "woman", "all persons" is judged to be more specific than "all women".

With judgments of concept specificity in place, it is possible to determine entailment relations on the basis of (un)instantiability claims in the T and H AKRs. For example, suppose the T and H AKRs contain the facts in (3).

(3) T: instantiable(C_T, Ctx_T)
H: instantiable(C_H, Ctx_H)

where concept C_T is aligned with C_H, C_T is judged to be more specific than C_H, and context Ctx_T is aligned with context Ctx_H. In this case, the hypothesis instantiability claim is entailed by the text instantiability claim (existence of something more specific entails existence of something more general). This being so, the H instantiability claim can be deleted without loss of information.

If instead we had the (un)instantiability claims in (4) for the same alignments and specificity relations,

(4) T: instantiable(C_T, Ctx_T)
H: uninstantiable(C_H, Ctx_H)

we would have a contradiction: the text says that there is something of the more specific type C_T, whereas the hypothesis says there are no things of the more general type C_H. In this case, the rules explicitly flag a contradiction.

Once all (un)instantiability claims have been compared, it is possible to judge whether the text entails or contradicts the hypothesis. Entailed hypothesis (un)instantiability assertions are deleted from the representation. Consequently, if there is one T and H AKR readings and one set of alignments under which all the H (un)instantiability assertions have been removed, then there is an entailment of H by T. If there is a pair of readings and a set of alignments under which a contradiction is flagged, then there is a contradiction. If there is no pair of readings or set of alignments under which there is either an entailment or a contradiction, then T and H are merely consistent with one another. There are exceptional cases such as (5) where one reading of T entails H and another reading contradicts it.

(5) T: John did not wait to call for help.
H: John called for help.

Our ECD rules detect such cases.

WordNet often misses synonyms needed for the alignment in the ECD. In particular, the hierarchy and synsets for verbs are one of WordNet's least developed parts. To test the impact of the missing synonyms, we developed a variation on the ECD algorithm that allows loose matching.

First, in concept alignment, if a verb concept in H does not align with any verb concept in T, then we permit it to (separately) align with all the text verb concepts. We do not permit the same loose alignment for noun concepts, since we judge WordNet information to be more reliable for nouns. This free alignment of verbs might sound risky, but in general these alignments will not lead to useful concept specificity judgments unless the T and H verbs have very similar arguments / role restrictions.

When such a loose verb alignment is made, we explicitly record this fact in a justification term included in the alignment fact. Similarly, when judging concept specificity, each rule that applies adds a term to a list of justifications recorded as part of the fact indicating the specificity relation. This means that when the final specificity judgments are determined, each judgment has a record of the sequence of decisions made to reach it.

(Un)instantiability comparisons are made as in strict matching. However, the criteria for detecting an entailment are selectively loosened. If no contradiction is flagged, and there is a pairing of readings and alignments under which just a single H instantiability assertion is left standing, then this is allowed through as a loose entailment. However, further rules are applied to block those loose entailments that are deemed inappropriate. These blocking rules look at the form of the justification terms gathered based on specificity judgments.

These blocking rules are manually selected. First, a loose matching run is made without any blocking rules. Results are dumped for each T-H pair, recording the expected logical relation and the justifications collected. Blocking rules are created by detecting patterns of justification that are associated with labeled non-entailments. One such blocking rule says that if you have just a single H instantiability left, but the specificity justifications leading to this have been shown to be reliable on training data, then the instantiability should not be eliminated as a loose entailment.

4 Challenges in Processing the RTE Data

The RTE3 data set contains inconsistencies in spelling and punctuation between the text and the hypothesis. To handle these, we did an automatic prepass where we compared the strings in the passage text to those in the hypothesis. Some of the special cases that we handled include:

- Normalize capitalization and spacing
- Identify acronyms and shorten names
- Title identification
- Spelling correction

Role names in VerbNet are in part intended to capture the relation of the argument to the event being described by the verb. For example, an object playing an Agent role is causally involved in the event, while an object playing a Theme or Patient role is only supposed to be affected. This allows participants in an action to be identified regardless of the syntactic frame chosen to represent the verb; this was seen in (1) and (2). Sometimes the roles from VerbNet are not assigned in such a way as to allow such transparent identification across frames or related verbs. Consider an example. In *Ed travels/goes to Boston* VerbNet identifies Ed as playing a Theme role. However, in *Ed flies to Boston* VerbNet assigns Ed an Agent role; this difference can make determining contradiction and entailment between T and H difficult. We have tried to compensate in our ECD, by using a backoff strategy where fewer role names are used (by projecting down role names to the smaller set). As we develop the system further, we continue to experiment with which set of roles works best for which tasks.

Another open issue involves identifying alternative ways vague relations among objects appear in text. We do not match the expression *the Boston team* with *the team from Boston*. To improve our recall, we are considering loose matching techniques.

5 Summary of our results on RTE3

We participated in the RTE challenge as a way to understand what our particular techniques could do with respect to a more general version of textual entailment. The overall experiment was quite enlightening. Tables 1 and 2 summarize how we did on the RTE3 challenge. System 1 is our standard system with strict ECD. System 2 used the looser set of ECD rules.

	Gold YES	Sys YES	Correct	R	P	F
IE	105	6	5	0.048	0.83	0.20
IR	87	4	4	0.046	1.00	0.21
QA	106	10	9	0.085	0.90	0.28
SUM	112	11	7	0.063	0.64	0.20
Total	410	31	25	0.060	0.84	0.22

Table 1: **System 1 with Strict ECD**

	Gold YES	Sys YES	Correct	R	P	F
IE	105	15	10	0.095	0.67	0.25
IR	87	6	4	0.046	0.67	0.18
QA	106	14	13	0.12	0.93	0.34
SUM	112	17	10	0.089	0.59	0.23
Total	410	52	37	0.088	0.71	0.25

Table 2: **System 2 with Loose** ECD

As can be seen, we answered very few of the questions; only 31 of the possible 410 with a YES answer. However, for those we did answer (requiring only linguistic, and not world knowledge), we achieved high precision: up to 90% on QA. However, we were not perfect even from this perspective. Here are simplified versions of the errors where our system answered YES, and the answer should be NO with an analysis of what is needed in the system to correct the error.

The wrong result in (6) is due to our incomplete coverage of intensional verbs (*seek*, *want*, *look for*, *need*, etc.).

(6) T: The US sought the release of hostages.
H: Hostages were released.

The object of an intensional verb cannot be assumed to exist or to occur. Intensional verbs need to be marked systematically in our lexicon.

The problem with (7) lies in the lack of treatment for generic sentences.

(7) T: Girls and boys are segregated in high school during sex education class.
H: Girls and boys are segregated in high school.

The natural interpretation of H is that girls and boys are segregated in high school ALL THE TIME. Because we do not yet handle generic sentences properly, our algorithm for calculating specificity produces the wrong result here. It judges segregation in H to be less specific than in T whereas the opposite is in fact the case. Adding the word "sometimes" to H would make our YES the correct answer.

The distinction between generic and episodic readings is difficult to make but crucial for the interpretation of bare plural noun phrases such as *girls* and *boys*. For example, the most likely interpretation of *Counselors are available* is episodic: SOME counselors are available. But *Experts are highly paid* is weighted towards a generic reading: MOST IF NOT ALL experts get a good salary.

These examples are indicative of the subtlety of analysis necessary for high precision textual inference.

References

Danny Bobrow, Cleo Condoravdi, Richard Crouch, Ronald Kaplan, Lauri Karttunen, Tracy Holloway King, Valeria de Paiva, and Annie Zaenen. 2005. A basic logic for textual inference. In *Proceedings of the AAAI Workshop on Inference for Textual Question Answering*.

Dick Crouch and Tracy Holloway King. 2005. Unifying lexical resources. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*.

Dick Crouch and Tracy Holloway King. 2006. Semantics via F-structure rewriting. In *Proceedings of LFG06*. CSLI On-line Publications.

Dick Crouch, Mary Dalrymple, Ron Kaplan, Tracy King, John Maxwell, and Paula Newman. 2007. XLE documentation. Available on-line.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Olga Gurevich, Richard Crouch, Tracy Holloway King, and Valeria de Paiva. 2006. Deverbal nouns in knowledge representation. In *Proceedings of FLAIRS 2006*.

Andres Hickl, John Williams, Jeremy Bensley, Kirk Roberts, Bryan Rink, and Ying Shi. 2006. Recognizing textual entailment with LCC's GROUNDHOG system. In *The Second PASCAL Recognising Textual Entailment Challenge*.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *AAAI-2000 17th National Conference on Artificial Intelligence*.

John Maxwell and Ron Kaplan. 1991. A method for disjunctive constraint satisfaction. *Current Issues in Parsing Technologies*.

Rowan Nairn, Cleo Condoravdi, and Lauri Karttunen. 2006. Computing relative polarity for textual inference. In *Proceedings of ICoS-5*.

Stefan Riezler, Tracy Holloway King, Ron Kaplan, Dick Crouch, John Maxwell, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.