Entailment and Anaphora Resolution in RTE3

Rodolfo Delmonte, Antonella Bristot, Marco Aldo Piccolino Boniforti, Sara Tonelli
Department of Language Sciences
Università Ca' Foscari – Ca' Bembo
30123, Venezia, Italy
delmont@unive.it

Abstract

We present VENSES, a linguistically-based approach for semantic inference which is built around a neat division of labour between two main components: a grammatically-driven subsystem which is responsible for the level of predicate-arguments well-formedness and works on the output of a deep parser that produces augmented head-dependency structures. A second subsystem fires allowed logical and lexical inferences on the basis of different types of structural transformations intended to produce a semantically valid meaning correspondence. In the current challenge, we produced a new version of the system, where we do away with grammatical relations and only use semantic roles to generate weighted scores. We also added a number of additional modules to cope with fine-grained inferential triggers which were not present in previous dataset. Different levels of argumenthood have been devised in order to cope with semantic uncertainty generated by nearly-inferrable Text-Hypothesis pairs where the interpretation needs reasoning. RTE3 has introduced texts of paragraph length: in turn this has prompted us to upgrade VENSES by the addition of a discourse level anaphora resolution module, which is paramount to allow entailment in pairs where the relevant portion of text contains pronominal expressions. We present the system, its relevance to the task at hand and an evaluation.

1 Introduction

The system for semantic evaluation VENSES (Venice Semantic Evaluation System) is organized as a pipeline of two subsystems: the first is a reduced version of GETARUN, our system for Text Understanding; the second is the semantic evaluator which was previously created for Summary and Question evaluation and has now been thoroughly revised for the new more comprehensive RTE task.

The reduced GETARUN is composed of the usual sequence of sub-modules common in Information Extraction systems, i.e. a tokenizer, a multiword and NE recognition module, a PoS tagger based on finite state automata; then a multilayered cascaded RTN-based parser which produces c-structures and has an additional module to map them into tentative grammatical functions in LFG terms. The functionally labeled constituents are then passed to an interpretation module that uses subcategorization information to choose final grammatical relations and assign semantic roles. Eventually, the system has a pronominal binding module that works at clause level for lexical personal, possessive and reflexive pronouns, which are substituted by the heads of their antecedents - if available. The output of the binding module can contain one or more "external" pronouns, which need to be bound in the discourse. This output is passed to the Anaphora Resolution module presented in detail in Delmonte (2006) and outlined below. This module works on the so-called History List of entities present in the text so far. In order to make the output of this module usable by the Semantic Evaluator, we decided to produce a flat list of semantic vectors which contain all semantic related items of the current sentence. Inside these vectors, pronominal expressions are substituted by the heads of their antecedents.

Basically, the output of the system is elaborated on top of the output of the parser, which produces a flat list of fully indexed augmented head-dependent structures (AHDS) with Grammatical Relations (GRs) and Semantic Roles (SRs) labels. Notable additions to the usual formalism is the presence of a distinguished Negation relation; we

Proceedings of the Workshop on Textual Entailment and Paraphrasing, pages 48–53,
Prague, June 2007.

also mark modals and progressive mood, tense and voice (for similar approaches see Bos et al., Raina et al.).

The evaluation system uses a cost model with rewards/penalties for T/H pairs where text entailment is interpreted in terms of semantic similarity: the closest the T/H pairs are in semantic terms the more probable is their entailment. Rewards in terms of scores are assigned for each "similar" semantic element; penalties on the contrary can be expressed in terms of scores or they can determine a local failure and a consequent FALSE decision – more on scoring below.

The evaluation system is made up of two main Modules: the first takes care of paraphrases and idiomatic expressions; the second is a sequence of linguistic rule-based sub-modules. Their work is basically a count of how much similar are linguistic elements in the H/T pair. Similarity may range from identical linguistic descriptions, to synonymous or just morphologically derivable ones. As to GRs and SRs, they are scored higher according to whether they belong to the subset of core relations and roles, i.e. obligatory arguments, or not, that is adjuncts. Both Modules go through General Consistency checks which are targeted to high level semantic attributes like presence of modality, negation, and opacity operators. The latter ones are expressed either by the presence of discourse markers of conditionality or by a secondary level relation intervening between the main predicate and a governing higher predicate belonging to the class of non factual verbs, but see below.

All rule-based sub-modules are organized into a sequence of syntactic-semantic transformation rules going from those containing axiomatic-like paraphrase HDSs which are ranked higher, to rules stating conditions for similarity according to a scale of argumentality (more below) and are ranked lower. All rules address HDSs and SRs. Propositional level ones have access to vectors of semantic features which encode propositional level information.

2 The Task of Semantic Inference Evaluation

As happened in the previous Challenge, this year's Challenge is also characterized by the presence of a relatively high number (we counted more than 100 True and another 100 False pairs, i.e. 25%) of T/H pairs which require two particularly hard NLP processes to set up:

- reasoning
- paraphrases (reformulation)

In addition to that, we found a significant number of pairs – about 150 in the development set and some 100 in the test set – in which pronominal binding and anaphora resolution are essential to the definition of entailment. In particular, in such cases it is only by virtue of the substitution of a pronominal expression with the linguistic description of its antecedent that the appropriate Predicate-Argument relations required in order to fire the inference is made available – but see below.

Setting up rules for Paraphrase evaluation, requires the system to actually use the lemmas – and in some cases the wordforms - to be matched together in axiomatic-like structures: in other words, in these cases the actual linguistic expressions play a determining role in the task to derive a semantic inference. In order for these rules to be applied by the SE, we surmise it is important to address a combination of Syntactic and Semantic structures, where Lexical Inference also may play a role: the parser in this case has a fundamental task of recovering the heads of antecedents of all possible pronominal and unexpressed Grammatical relations. It is important to remark that besides pronominal-antecedent relations, our system also recovers all those cases defined as Control in LFG theory, where basically the unexpressed subject of an untensed proposition (infinitival, participial, gerundive) either lexically, syntactically or structurally controlled is bound to some argument of the governing predicate.

2.1 Pronominal Binding and Anaphora Resolution

This year RTE introduces as a novelty a certain number (117 in the Test set – 135 in the Dev set) of long Texts, of paragraph length. This move is justified by the need to address more realistic data, and consequently to tune the whole process of semantic evaluation to the problems related to such data. Thus more relevance is given to empirical issues to be tackled, rather than to the theoretical ones, which however don't disappear but may assume less importance.

When a system has to cope with paragraph length texts, the basic difference with short texts regards the problem of anaphora resolution. In short texts, pronominal expressions constituted a minor problem and all referring expressions were specified fully. Not so in long texts, as can be seen from the Table below:

	He	Him	His	She	Her	It	Its	They	Their	Them	Total
Test	80	15	91	19	18	91	68	43	63	15	485
Dev.	113	16	136	27	35	123	76	44	64	18	652
Total	193	31	227	46	53	214	144	87	127	33	1137

Table 1. 3rd person pronominal expressions contained in RTE3 data sets

As can be seen from this table, the problem a system is faced with is not just to cope with an ad hoc solution for single cases where the pronoun is placed, for instance, in sentence first position and it might be easy to recover its antecedent by some empirical ad hoc procedure. The problem needs to be addressed fully and this requires a full-fledged system for anaphora resolution. One such system is shown in Fig. 2 below, where we highlight the architecture and main processes undergoing at the anaphora level. First of all, the subdivision of the system into two levels: Clause level – intrasentential pronominal phenomena – where all pronominal expressions contained in modifiers, adjuncts or complement clauses receive their antecedent locally. Possessive pronouns, pronouns contained in relative clauses and complement clauses choose preferentially their antecedents from list of higher level referring expressions. Not so for those pronouns contained in matrix clauses. In particular the ones in subject position are to be coreferred in the discourse. This requires the system to be equipped with a History List of all referring expressions to be used when needed.

In the system, three levels are indicated: Clause level, i.e. simple sentences; Utterance level, i.e. complex sentences; Discourse level, i.e. intersententially. Our system computes semantic structures in a sentence by sentence fashion and any information useful to carry out anaphoric processes needs to be made available to the following portion of text, and eventually to the Semantic Evaluation that computes entailment. We will comment a number of significant examples to clarify the way in which our system operates.

3. Anaphora Resolution for RTE

Why is it important to implement an anaphora resolution algorithm for RTE? I think the reason is quite straightforward: pronominal expressions do not allow any inference to be drawn or any otherwise semantic similarity processes to be fired, being inherently referentially poor. In order to be able to use information made available by the verb in the sentence in which a pronoun is used, the antecedent must be recovered and the pronoun substituted by its head. So, very simply, RTE needs anaphora resolution in order to allow the system to use verb predicates where pronouns have been used to corefer to some antecedent in the previous text. In turn that verb predicate is just what is being searched for in the first place, and in our case it is the one contained in the Hypothesis.

The current algorithm for anaphora resolution works on the output of the complete deep robust parser which builds an indexed linear list of dependency structures where clause boundaries are clearly indicated. As said above, our system elaborates both grammatical relations and semantic roles information for arguments and adjuncts. Semantic roles are very important in the weighting procedures.

As to the anaphoric resolution algorithm, it is a distributed, local – clause-based - approach to anaphora resolution which we regard more efficient than monolithic, global ones. Linguistic theory has long since established without any doubt the existence in most languages of the world of at least two classes of pronouns: the class which must be bound locally in a given domain – roughly the clause, and the class which must be left free in the same domain.

In our approach, we proceed in a clause by clause fashion, weighting each candidate antecedent w.r.t. that domain, trying to resolve it locally. Weighting criteria are amenable on the one hand to linear precedence constraints, with scores assigned on a functional/semantic basis. On the other hand, these criteria may be overrun by a functional ranking of clauses which requires to treat main clauses differently from secondary clauses, and these two differently from complement clauses.

There are also two general referential policy assumption that we adopt in our approach: The first one is related to pronominal expressions, the second one to referring expressions or entities to be asserted in the History List, and are expressed as follows:

- no more than two pronominal expressions are allowed to refer back in the previous discourse portion;
- at discourse level, referring expressions are stored in a push-down stack according to Persistence principles.

Only "persistent" referring expressions are allowed to build up the History List, where persistence is established on the basis of the frequency of topicality for each referring expression which must be higher than 1. All referring expression asserted as Topic (Main, Secondary, Potential) only once are discarded in case they appeared at a distance measured in 5 previous utterances. Proximate referring expressions are allowed to be asserted in the History List. The first procedure is organized as follows.

A. For each clause,

1. we collect all referential expressions and weight them – this is followed by an automatic ranking;
2. then we subtract pronominal expressions;
3. at clause level, we try to bind personal and possessive pronouns obeying specific structural properties; we also bind reflexive pronouns and reciprocals if any, which must be bound obligatorily in this domain;
4. when binding a pronoun, we check for disjointness w.r.t. a previously bound pronoun if any;
5. all unbound pronouns and all remaining personal pronouns are asserted as "externals", and are passed up to the higher clause levels;

B. Then we turn at the higher level – if any -, and we proceed as in A., in addition we try to bind pronouns passed up by the lower clause levels

- o if successful, this will activate a retract of the "external" label and a label of "antecedenthood" for the current pronoun with a given antecedent;
- o the best antecedent is chosen by recursively trying to match features of the pronoun with the first available antecedent previously ranked by weighting;
- o here again whenever a pronoun is bound we check for disjointness at utterance level.

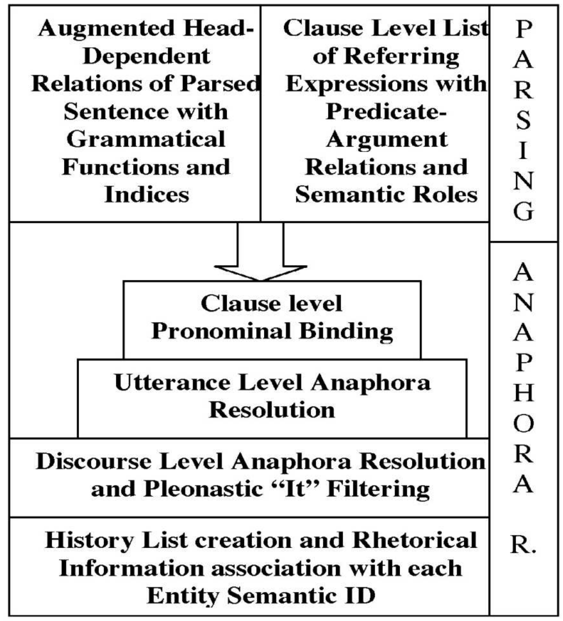

Fig. 1. Anaphoric Processes in VENSES

C. This is repeated until all clauses are examined and all pronouns are scrutinised and bound or left free.

D. Pronouns left free – those asserted as externals – will be matched tentatively with the best candidates provided this time by a "centering-like" algorithm. Step A. is identical and is recursively repeated until all clauses are processed.

3.1 Focussing Revisited

Our version of the focussing algorithm follows Sidner's proposal (Sidner C., 1983; Grosz B., Sidner C., 1986), to use a Focus Stack, a certain Focus Algorithm with Focus movements and data structures to allow for processing simple inferential relations between different linguistic descriptions co-specifying or coreferring to a given entity.

Our Focus Algorithm is organized as follows: for each utterance, we assert three hierarchically ordered "centers" that we call Main, Secondary and the first Potential Topic, which represent the best three referring expressions as they have been weighted in the candidate list used for pronominal binding; then we also keep a list of Potential Topics for the remaining best candidates. These three best candidates repositories are renovated at each new utterance, and are used both to resolve pronominal and nominal cospecification and coreference: this is done both in case of strict identity of linguistic description and of non-identity.

The Main Topic may be regarded the Forward Looking Center in the centering terminology or the Current Focus. All entities are stored in the History List (HL) which is a stack containing their morphological and semantic features. In the HL every new entity is assigned a semantic index which identifies it uniquely. To allow for Persistence evaluation, we also assert rhetorical properties associated to each entity, i.e. we store the information of topicality (i.e. whether it has been evaluated as Main, Secondary or Potential Topic), together with the semantic ID and the number of the current utterance. This is subsequently used to measure the degree of Persistence in the overall text of a given entity, as explained below.

In order to decide which entity has to become Main, Secondary or Potential Topic we proceed as follows:

- we collect all entities present in the History List with their semantic identifier and feature list and proceed to an additional weighting procedure;
- nominal expressions, they are divided up into four semantic types: definite, indefinite, bare NPs, quantified NPs. Both definite and indefinite NP may be computed as new or old entity according to contextual conditions as will be discussed below and are given a rewarding score;
- we enumerate for each entity its persistence in the previous text, and keep entities which have frequency higher than 1, we discard the others;
- we recover entities which have been asserted in the HL in proximity to the current utterance, up to four utterances back;
- we use this list to "resolve" referring expressions contained in the current utterance;
- if this succeeds, we use the "resolved" entities as new Main, Secondary, and Potential Topics and assert the rest in the Potential Topics stack;
- if this fails – also partially – we use the best candidates in the weighted list of referring expressions to assert the new Topics. It may be the case that both resolved and current best candidates are used, and this is by far the most common case.

In example n.3 below, the first possessive pronoun "his" is met at Utterance level – the first sentence has two clauses: clause 1, headed by the predicate DIVORCE, and clause 2, headed by MARRY. "His" will look for a masculine antecedent and Chabrol will be chosen, also for weights associated to it, being the higher subject. This will produce the following semantic structure, which is made of a Head, a Semantic Role and an Index,

- Chabrol-poss-sn2

which is the output of the substitution of "his" present in the same structure by means of information made available by the Anaphoric module. Note that the index of a modifier points to the governing head, in this case "wife", the apposition associated to "Agnes", which in turn is the OBJect of DIVORCE.

T/H pair n. 3
T: Claude Chabrol divorced Agnes, his first wife, to marry the actress Stèphane Audran. His third wife is Aurore Paquiss.
H: Aurore Paquiss married Chabrol.

When the first sentence is passed to the semantic interpreter, anaphoric processes have already been completed and the information is then transferred to semantic structure which will register the anaphoric relation by the substitution operation. However this specific relation is not the one that really matters in the current T/H pair. When the system passes to the analysis of the following sentence it has another possessive pronoun which is contained in a SUBJect NP. By definition, these pronouns take their antecedent from the discourse level. To have the system do that, the pronoun has to be left free at sentence level, i.e. it must be computed as "external" to the current sentence, and not bound locally. Discourse level processes will look for antecedents from the History list and from the socalled Topic Hierarchy, our way to compute centering (but see again Delmonte, 2006). This is shown schematically in the output of the Anaphora Resolution module shown here below, which reports the listing of pronouns, Topic Hierarchy, and Anaphora Resolution processes carried out. In this case, every referring expression will have a semantic index (SI) associated which is unique in the History List. In example n.31, here below, the pronominal expressions are two: an Utterance level possessive pronoun bound to the local SUBJect; and a Discourse level personal pronoun “He” which receives its antecedent from the History List. In both cases, substitution with their antecedents’ head will take place in the semantic interpretation level.

T/H pair n. 31
T: Upon receiving his Ph.D., Wetherill became a staff member at Carnegie's Department of Terrestrial Magnetism (DTM) in Washington, D.C. He originated the concept of the Concordia Diagram for the uranium-lead isotopic system.
H: Wetherill was the inventor of the concept of the Concordia Diagram.

Clause No.	Main Topic	Secondary Topic	Potential Topics	Pronouns + Features	Disjoint-ness	Pronominal Binding	Anaphora Resolution
id_3_1 Text	'Claude Chabrol'		Agnes				
id_3_2 Text	**Main resolved as** Claude Chabrol - SI=id1	wife	Aurore Paquiss	his-[sem:hum, cat:poss, gen:mas, num:sing, pers:3, pred:he, gov_pred:be]	disj=[sn1-wife]	**External pronoun** (be, his)	his **resolved as** 'Claude Chabrol'
id_3_3 Hypo-thesis	Aurore Paquiss - SI=id4	Claude Chabrol - SI=id1					

Table 2. Output of the Anaphora Resolution Module

4 Results and Discussion

Results for the Test set 485 total pronominal expressions amount to 69% accuracy, 92% recall – this includes computing It-expletives. The F-measure computed is thus 79%. Overall, we evaluated the contribution of the Anaphora Resolution Module as 15% additional correct results. Of course, the impact of using this module would have been different in case all T/H pairs were constituted by long texts.
The RTE task is a hard task: in our case 10-15% mistakes are ascribable to the parser or any other analysis tool; another 5-10% mistakes will certainly come from insufficient semantic information.

	ACCURACY	AVERAGE PRECISION
IE	**0.5850**	**0.5992**
IR	**0.6050**	**0.5296**
QA	**0.5900**	**0.6327**
SUMM	**0.5700**	**0.6132**
TOTAL	**0.5875**	**0.5830**

Table 3. Results for First Run

References

Bos, J., Clark, S., Steedman, M., Curran, J., Hockenmaier, J.: Wide-coverage semantic representations from a ccg parser. In: Proc. of the 20th International Conference on Computational Linguistics. Geneva, Switzerland (2004)

Delmonte, R.: Text Understanding with GETARUNS for Q/A and Summarization, Proc. ACL 2004 - 2nd Workshop on Text Meaning & Interpretation, Barcelona, Columbia University (2004) 97-104

Delmonte R., et al. Another Evaluation of Anaphora Resolution Algorithms and a Comparison with GETARUNS' Knowledge Rich Approach. In: ROMAND 2006 - 11th EACL. Geneva, (2006) 3-10

Grosz B. and C. Sidner 1986. Attention, Intentions, and the Structure of Discourse, *Computational Linguistics* 12 (3), 175-204.

Raina, R., et al.: Robust Textual Inference using Diverse Knowledge Sources. In: Proc. of the 1st. PASCAL Recognision Textual Entailment Challenge Workshop, Southampton, U.K., (2005) 57-60

Sidner C. 1983. Focusing in the Comprehension of Definite Anaphora, in Brady M., Berwick R.(eds.), *Computational Models of Discourse*, MIT Press, Cambridge, MA, 267-330.