

VENSES – a Linguistically-Based System for Semantic Evaluation

**Rodolfo Delmonte, Sara Tonelli, Marco
Aldo Piccolino Boniforti, Antonella Bristot**

Department of Language Sciences
University Ca' Foscari
30124 Venice, Italy
delmont@unive.it

Emanuele Pianta

ITC-IRST
POVO - TRENTO

pianta@itc.it

Abstract

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 output of the system is a flat list of head-dependent structures (HDS) with Grammatical Relations (GRs) and Semantic Roles (SRs) labels. The evaluation system is made up of two main modules: the first is a sequence of linguistic rule-based subcalls; the second is a quantitatively based measurement of input structures. VENSES measures semantic similarity which may range from identical linguistic items, to synonymous or just morphologically derivable. Both modules go through General Consistency checks which are targeted to high level semantic attributes like presence of modality, negation, and opacity operators, temporal and spatial location checks.

Results in cws, accuracy and precision are homogenous for both training and test corpus and fare higher than 60%.

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 submodules 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 is equipped with an interpretation module that uses subcategorization information and semantic roles processing. Eventually, the system is equipped with a pronominal binding module that works for lexical personal, possessive and reflexive pronouns, which are substituted by the heads of their antecedents - if available. The output of the system is

a flat list of head-dependent structures (HDS) with Grammatical Relations (GRs) and Semantic Roles (SRs) labels. Notable additions to the usual formalism is the presence of a distinguished Negation relation; we also mark modals and progressive mood. All other non semantic elements like auxiliaries and determiners are erased.

The evaluation system uses a strategy of 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.

The evaluation system accesses the output of GETARUN which sits on files and is totally independent of it. It is made up of two main Modules: the first is a sequence of linguistic rule-based subcalls; the second is a quantitatively based measurement of input structures. The latter is basically a count of heads, dependents, GRs and SRs, scoring only similar elements in the H/T pair. Similarity may range from identical linguistic items, to synonymous or just morphologically derivable. 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 as 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. Two other general consistency calls regard temporal and spatial location checks which must be identical or entailed in one another, if present – but see below.

Linguistic rule-based subcalls are organized into a sequence of calls going from rules containing axiomatic-like paraphrase HDSs which are ranked higher, to rules stating conditions for similarity according to the scale of argumentality which are ranked lower. All rules address HDSs, GRs and SRs. Both Modules strive for True assessments: however, Calls 1 are then followed by Calls 2 which can

output True or False according to general consistency or scoring. Modifying the scoring function may thus vary the final result dramatically: it may contribute more True decisions if relaxed, so it needs fine tuning. More experimentation is needed on much bigger data set to achieve a more general definition of this function.

2. An A-As Hybrid Parser

Our parser has been presented in detail lately in a number of papers and has achieved 90% recall on Greval Corpus and 89% recall on the XEROX-700 corpus, limited only this latter test to SUBJ/OBJ GRs. As in most robust parsers, we use a sequence or cascade of transducers: however, in our approach, since we intend to recover sentence level structure, the process goes from partial parses to full sentence parses. Sentence and then clause level parsing are crucially responsible for the right assignment of Arguments and Adjuncts (hence A-As) to a governing predicate head. This is paramount in our scheme which aims at recovering predicate-argument structures, besides performing a compositional semantic translation of each semantically headed constituent.

The first transducer receives the input sentence split by previous processors, which is recursively/iteratively turned into a set of non-sentential level syntactic constituents - some of which can incorporate a PP headed by "of". Non-sentential level constituents, can be interspersed by heads which are subordinate clause markers, like subordinating conjunctions, or parentheticals - by punctuation, indirect interrogative clauses - by interrogative pronouns. The final output is a list of headed syntactic constituents which comprise the usual set of semantically translatable constituents, i.e., ADJP, ADVP, NP, PP, VC (Verb Cluster).

The task of the following transducer is that of creating clauses: we assume that at each sentence level only one VCluster (hence VC) can appear: we define the VC as IBAR indicating that there must be a finite or tensed verb included in it. VCs containing non-tensed verbal elements are all defined separately.

The third pass is intended to produce an improvement on the sentence-level full parse, by transducing each constituent label into a corresponding grammatical function label. The rules are taken from the inventory of LFG theory and follow its rules and principles. All attachment decisions are taken at this level of computation. In particular both PP and Relative Clauses are attached locally according to preferences and best match (but see Delmonte 2002). Finally the fourth pass has the task of splitting complex sentences into simplex ones, or clauses.

The output of the four transducers is passed to the algorithm that takes care of the creation of predicate-

argument structures which has the additional task of taking into due account interclausal relations. To do that, semantic indices of governing predicates are used to assert dependencies between two adjacent clauses. This may also apply to a main clause and a clause-like adjunct like a gerundive or a participial. Lexical information is accessed to confirm or modify previous decisions, particularly as regards OBLiques which will be interpreted as Adjuncts or Argument at this level of interpretation. We also assert Semantic Roles on the basis of lexical information (see Delmonte 1990).

To be compliant with usual Dependency Structure inventory of GRs which we also had to use for evaluation purposes, we eventually turn all predicative labels - NCOMP, ACOMP, PCOMP, VCOMP - into XCOMP. Also OBLiques are turned into IOBJect, unless they represent the passive agent by-adjunct which is assigned the GR label ARG_MOD. Then we produce flat Head-Dependent Structures.

We don't have space here to describe the Pronominal Binding module which however accesses Referential Heads at clause level and establishes possible antecedent-pronoun candidate lists which are then weighted and the best one chosen (but see Delmonte, Bianchi 1991). As an example consider Snippet 820 reported here below:

T. Clinton's new book is not big seller here.
H. Clinton's book is a big seller.

Whose structure is computed respectively as follows:

T.
be(adj-locative, here).
seller(ncmod, big).
book(ncmod-specif, 'Clinton-s_').
be(xcomp-prop, seller).
be(subj- theme_bound, book).
be(neg, not).
H.
seller(ncmod, big).
book(ncmod-specif, 'Clinton-s_').
be(xcomp-prop, seller).
be(subj-theme_bound, book).

The presence of the negation operator in the T portion of the snippet will prevent the evaluator from assessing to TRUE even though the relevant HD structures are identical.

3. The Semantic Evaluator (SE)

As said above, the SE is organized into two main modules: a quantitatively based module, and a sequence of rule-based subcalls where scoring is also taken into account when needed, to increase confidence in the decision process. The two modules must then undergo general consistency checks which have the task to ascertain the presence of possible

mismatches at semantic level. In particular, these checks take care of the following semantic items:

- presence of spatiotemporal locations relatively to the same governing predicate, or a similar one as has been computed from previous modules;
- presence of opacity operators like discourse markers for conditionality having scope over the governing predicate under analysis;
- presence of quantifiers and other referentiality related determiners attached to the same nominal head in the T/H pair under analysis and chosen as relevant one by previous computation;
- presence of antonyms in the T/H pair at the level of governing predicates;
- presence of predicates belonging to the class of “doubt” expressing verbs, governing the relevant predicate shared by the T/H pair.

In some cases the General Consistency Checks have to be suspended: in particular whenever both T/H pairs contain opacity operators and negation, as for instance in,

Snippet no. 1014

The thick atmosphere of Titan makes it difficult for even the largest telescopes on Earth to see anything clearly.

Telescopes on Earth cannot see Titan clearly.

3.1 The Rule-Based Module

This Module is organized as a sequence of rule-based calls which start from exceptional cases down to default cases. Exceptional cases of Semantic Similarity are those constituted by definition-like H sentences, or simple paraphrases of the meaning expressed by the main predicate of the T text. Generally speaking, every time one such rule is fired, the T/H pair contains a conceptually complex lexical predicate and its paraphrase in conceptually simple components.

Examples of such cases are constituted by pairs like the following:

- a. interview --> conduct an interview
- b. pressurise --> apply pressure
- c. treat --> receive treatment (provide)
- d. fire → send letter of dismissal

where both a. and b. were actually present in WordNet while c. did not figure with the same predicates but rather with the one in brackets; d. was totally absent.

Definitions and paraphrases are looked up at first in the definitions made available by WordNet. In case of failure a list of some 50 manually made up axiomatic rules are accessed. Each such rule addresses main predicates in the T/H pair, together with presence of semantically relevant dependent if needed, and whenever the concept expressed by the lexically complex predicate requires it. Together with the predicates, the rules select relevant GRs and

SRs when needed. In addition, more restrictions are introduced on additional arguments or adjuncts. Eventually, as is the case with all the rules, penalties are explored in terms of semantic operators of the main predicate like negation, modality and opacity inducing verbs which must either absent or be identical in the T/H pair.

The linguistically-based Module is organized into a sequence of five subcalls where the T/H pairs are checked for semantic similarity starting from sameness of main predicates to semantic approximate match.

The first subcall requires the presence of same HDs as main predicates with core arguments, i.e. the ones which have been computed as subject, object, indirect object, arg_mod (passive “by” agent adjunct), xcomp. Nonconflicting SRs are checked in all subcalls: i.e. subject-agent are allowed to match with arg_mod-agent and subject-theme_affected with object-theme_affected but not viceversa. These matches take care of what are usually referred to as lexical alternations for verb categorization frames, and lexical rules in LFG terms which encompass such syntactic phenomena as passive, intransitivization, ergativization, dative shift, etc.

The second subcall requires the presence of same HDs as a combination of main head and main dependent and at least another identical HD structure within the core argument subset. Other subcalls included in this group check nominalization derivational relations intervening between main predicate of T and H, which in one case is checked with edit distance measures.

The third subcall takes as input a list of “light-verbs” in semantic terms, i.e. verbs including “be”, “have”, “appear”, and other similar copulative and locational verbs – like “live”, “hold”, “take_place”, “participate”, etc. - which are used to either make a definition, assert a property of the subject, individuate a location of the subject etc. These verbs are matched against main predicates and core arguments of the T portion, which must be identical to H. Quantitative measures are added to confirm the choice. Notable exceptions are sentences containing “be_born” predication which require specific constructions on the other member of the T/H pair.

The fourth subcall takes as input at least one identical main predicate HD non argument structure and one additional core argument or adjunct structure. Quantitative measures are added to confirm the choice.

The fifth subcall looks for different main predicates with core arguments which however must be non antonyms, non negative polarity and be synonyms. In addition, there must be at least another important identical non argument HD structure shared. Quantitative measures are added to confirm the choice. One such case is represented by

Snippet no. 1639

Lennon was murdered by Mark David Chapman outside the Dakota on Dec. 8, 1980.
Mark David Chapman killed Lennon.

Differently from what happens in real opposite meaning snippets where the SE considers SRs which must also be opposite, as in snippet 933,
Crude Oil Prices Slump
Oil prices drop

Or cases in which the snippet is rescued due to the presence of same SRs,
Snippet no. 876
Officials said Michael Hamilton was killed when gunmen opened fire and exchanged shots with Saudi security forces yesterday
Michael Hamilton died yesterday.

where DIE and KILL have opposite meaning but when KILL is used in the passive the SRs attached to their SUBJECTS will be identical.

3.2 The Quantitative Module

In this module all Heads, Dependents, GRs and SRs are collected for each member of the T/H pair and then they are passed to a scoring function that takes care of identical or similar members by assigning a certain score to every hit. Penalties correspond to high scores, while rewards correspond to low scores. A threshold is then set at a certain value which should encode the presence of a comparatively high number of identical/similar linguistic items.

As with previous subcalls, at the end of the computation semantic consistency and integrity is checked by collecting and comparing semantic operators, as well as performing a search of possible governing “doubt” verbs.

Generally speaking, we also treat short utterances differently from long ones. A stricter check is performed whenever an utterance has 3 or less HD structures, the reason being that in these structures some of the above mentioned subcalls would fail due to insufficient information available.

4. Evaluation and Discussion

The RTE task is a hard task: this may be partly due to the way in which it has been formulated – half of the snippets are TRUE, the other half are FALSE. It is usually the case that 10-15% mistakes are ascribable to the parser or any other analysis tool; another 5-10% mistakes will certainly come from insufficient semantic information. Whenever a system makes 20% errors this is doubled to 40% and the final result will become 60% overall Recall.

We looked into our mistakes to evaluate the import of the parser on the final Recall and we found out that: 10 snippets out of 100 TRUE ones have a wrong parse which can be regarded the main cause of the mistake. In other words only 10% of wrong

results can be ascribed to bad parses. The remaining 10% is due to insufficient semantic information. In turn, this may be classified as follows:

- 80% is due to lack of paraphrases and definitions;
- 10% is due to wrong SemanticRole assignment;
- 10% is due to lack of synonym/antonym relations.

When we started working on the training corpus, verb predicates synsets made available by WordNet have been augmented by the information contained in Grady Ward’s MOBY Thesaurus (<http://www.dcs.shef.ac.uk/research/ilash/Moby/>). Additional information has been derived from a manually reorganized version of Roget’s Thesaurus, again limited though to verb predicates. We also felt we needed information related to negative polarity verb predicates which we derived from Harvard Dictionary derived from Harvard IV-4 e Laswell's dictionary on the Dynamics of Culture (<http://www.wjh.harvard.edu/>). The paraphrase and definition list for verb predicates taken from WordNet and transformed into HD structures was also updated in order to cover some missing cases. For instance, we had to implement a new paraphrase for the verb FIRE which is paraphrased as “send dismissal letter to” in snippet no. 783. The list of HDSs will be accessed by the Evaluator in the appropriate Module.

Test-set Results	Training-set Results
cws: 0.6257	cws: 0.6396
accuracy: 0.5925	accuracy: 0.6032
precision: 0.6242	precision: 0.6261
recall: 0.4650	recall: 0.5088
f: 0.5330	f: 0.5614
CD cws: 0.7395 acc: 0.6867	CD cws: 0.7416 acc: 0.6633
QA cws: 0.5441 acc: 0.5846	QA cws: 0.5719 acc: 0.5444
PP cws: 0.8354 acc: 0.8000	PP cws: 0.6846 acc: 0.6707
IE cws: 0.6150 acc: 0.5833	IE cws: 0.6192 acc: 0.6000
IR cws: 0.6624 acc: 0.6222	IR cws: 0.6749 acc: 0.6286
RC cws: 0.5629 acc: 0.5214	RC cws: 0.5422 acc: 0.5243
MT cws: 0.4723 acc: 0.4667	MT cws: 0.6482 acc: 0.6111

Tab.1 Results for training and test-set

5. References

- Delmonte, R. 2003. Getaruns: a Hybrid System for Summarization and Question Answering. In Proc. Natural Language Processing (NLP) for Question-Answering, EACL, Budapest, ACL Columbia University, pp.21-28.
- Delmonte R. 2002. GETARUN PARSER - A parser equipped with Quantifier Raising and Anaphoric Binding based on LFG, Proc. LFG2002 Conference, Athens, pp.130-153, at <http://csli-publications.stanford.edu/hand/miscpubsonline.html>
- Delmonte, R. & D.Bianchi. 1991. Binding Pronominals with an LFG Parser, Proc. 2nd IWPT, Cancun(Messico), ACL 1991, pp. 59-72.