VENSESEVAL at Semeval-2016 Task 2: iSTS - with a full-fledged rule-based approach

Rodolfo Delmonte

Department of Language Studies and Comparative Cultures &
Department of Computer Science - Ca' Foscari University of Venice
Ca' Bembo – Dorsoduro 1075 - 30123 VENEZIA
Email:delmont@unive.it - Website: http//project.cgm.unive.it

Abstract

In our paper we present our rule-based system for semantic processing. In particular we show examples and solutions that may be challenge our approach. We then discuss problems and shortcomings of Task 2 – iSTS. We comment on the existence of a tension between the inherent need to on the one side, to make the task as much as possible "semantically feasible". Whereas the detailed presentation and some notes in the guidelines refer to inferential processes, paraphrases and the use of commonsense knowledge of the world for the interpretation to work. We then present results and some conclusions.

1 Introduction

In this presentation we will focus on Task 2, Interpretable Semantic Textual Similarity. We will comment on the way in which the task has been defined and what is eventually made available to participants to the task. In the section below we will present our system created for RTE – Recognizing Textual Entailment - challenges, and how it has been reorganized to suit the new current task. We discuss examples that suit or challenge our approach. We then comment in detail problems and shortcoming related to issues related to the annotations and the choice of semantic relations. In a final section we will present our results and a discussion.

2 The system *VENSESEVAL*

The system is an adaptation of *VENSES* Venice team system used for RTE (Recognizing Text Entailment) challenges (see Delmonte et al. 2009). In fact Venses was only partially adaptable to the new task, and so we had to partly recast it. In particular, the semantic evaluation module which was used to issue a binary (or ternary) decision in RTE challenges, in iSTS scenario works in a totally different manner. RTE required a Text - which could be constituted by a single simple or complex sentence - to be compared to an Hypothesis, this usually a simple sentence. From a purely theoretic and abstract point of view, measuring semantic similarity between two sentences (snippets) resembles very closely RTE as indicated in the paper accompanying 2015 challenge, (see Agirre et al. 2015). As the authors comment, iSTS can be defined as a graded similarity notion. In iSTS the comparison is between chunks which must be aligned first.

We find this approach certainly very useful and adequate for the task of semantic similarity checking. However, we had to reorganize and partly rewrite the modules for semantic matching. This has taken a lot of time and in fact the module hasn't been fully completed. So we assume that next year challenge will see a better – or at least complete - version of *VENSESEVAL*.

In the current challenge, we used only part of the original RTE system for various reasons. *Venses* was created to allow for semantic matching at different levels of complexity and representations. In particular, the task allowed semantic composition over different sentences, not necessarily adjacent to one another. It required anaphoric processes to be part of the semantic interpretation, again over a span of text made up of a number of different but referentially related sentences. For that reason, we used a complete system for semantic interpretation that not only had a level of anaphora and coreference resolution, it also ended up by creating a logical form representation which was then used for deep

Proceedings of SemEval-2016, pages 713–719,
San Diego, California, June 16-17, 2016.

semantic matching[1].

Venseseval has a different structure and less components. iSTS task seems much more a word or chunk level interpretation process. It does not require anaphora/coreference resolution because it is bound at sentence level. It does not require a logical form to be computed, or at least we haven't found a real motivation to do that. In fact, we then found good reasons for computing predicate-argument structures, but only after the deadline expired – more on this below. Eventually we decided to limit ourselves to produce syntactic constituency structure which could then be used to produce non-embedded chunks. The pipeline we used is then fairly simple:

- a tokenizer + a sentence splitter
- a tagger + an augmented recursive transition network
- a chunker extracting simple chunks - in the same fashion in which they have been characterized in the task - from sentence level constituency.

The system had to be reorganized around the gold chunk option which is what we wanted to experiment with. In order to allow chunk alignment to work correctly, sentences have been tokenized without introducing any further computation which could modify the order of the tokens. We just wanted to use tagged tokens with the gold chunks options, and chunked structures with the system option of the task.

In particular, in our original system, NP chunks were organized internally as a list to show the Head of the chunk in first position: all modifiers were moved to the right of the head. This was no longer possible, given the fact that we had to preserve the word order of the input sentence. No multiwords have been created, and this is something that may have contributed to decrease our tagger accuracy and as a consequence also chunks have a lower level of accuracy. The final system used the Named Entity Recognition module which however was only activated at chunk level, all other cases have been neglected. The reason for not producing multiwords was also related to the fact that we wanted to keep matching procedures as simple as possible. But we intend to reintroduce it in a future reimplementation of the system.

The first pass into the sentence pair is done with the aim to produce a general similarity evaluation based on tagged words only. To do this, we use part of the matching procedures explained below which erase function or stop words, and evaluate identity and similarity between content words without however issuing any score. Then we pass to the chunk-based analysis which is similar to the algorithm proposed in the Annotation Guidelines that can be taken as a starting point also for our module. For instance, as specified in the Procedure at pag.11, 3.d where it says "proceed from strongest to weakest 1:1 alignments". This is exactly what our system does, as specified in the algorithm below:

```
- given one sentence pair,
   - select first chunk in sentence one
      - recursively try to match chunk one to each of the chunks in sentence two
         - start matching procedures from EQUIvalence/OPPOsite
            - succeed, move to second chunk in sentence one
         - repeat
         - fail
         - move to matching procedures for SPE1/SPE2
            - succeed, move to second chunk in sentence one
         - fail
         - move to matching procedures for SIMI, REL
            - succeed, move to second chunk in sentence one
         - fail
         - assign NOALI label to chunk
         - move to second chunk in sentence one
         - repeat until end of chunks in sentence one
      - end
```

The main algorithm is a depth first attempt at finding the best match: we always select the first candidate available at each similarity level. We don't try all possible matches, in a breadth-first approach where we should score them and then choose the best candidate. At the end of the main recursion, the algorithm looks for possible recovery actions, by collecting all NOALI marked chunks and searching for possible matches with the already matched chunks except for the ones computed as EQUI. This is done trying to match

[1] A version of the RTE system is still working on our website, http://project.cgm.unive.it/cgi-bin/venses/venses.pl

the head of the NOALI chunk with the head of the companion chunk. In case of match, it substitutes the other chunk with the current one and turns it into a NOALI.

In the Annotation Guidelines we find two secondary features - Polarity and Factuality, appearing as extensions to main features, which are scarcely commented. We didn't implement these extensions because there were no clear instructions to do so in the guidelines. In the case of POLarity we only found four cases so labeled in the training gold standard for headlines – none in the images gold standard. As to FACTuality extensions we only implemented NOALI_FACT for all those cases in which a communication verb was labeled. Also in this case, however, there was no clear definition of the task in the guidelines, but we found 29 labeled cases in the headlines gold standard, none in the images. 13 examples were cases of NOALI_FACT which gave us sufficient confidence in assigning the label. The remaining cases were split between EQUI and SIMI with two SPE1 cases, and it was fairly hard to establish a rule that could fit with them all. SPE cases are individuated by matching the head of the two chunks which must be equal in the sense provided by EQUIvalence algorithm.

SIMI on the contrary requires some inferential step – but see section below. In particular we used a specialized label SIMI-2 to classify all chunk-pairs which involved differences in numbers, simple integers but also dates. Integers were then measured to check the distance in value and decide a score to associate to SIMI, which could vary from, 2 to 4. We used it also to indicate difference in country names, i.e. Locations recognized by the NER algorithm.

Problems arise for all those cases of paraphrases included in the corpus. There's a few examples in the training corpus of Headlines, in sentence 214,

//Iran says it captures drone ; U.S. denies losing one.
//Iran says it has seized U.S. drone ; U.S. says it 's not true
8 <==> 10 12 13 14 // EQUI // 5 // denies <==> says 's not true

Another such case is present in sentence 402,

//Egypt 's main opposition rejects president 's call for dialogue.
//Egypt opposition mulls response to Morsi dialogue call.
5 <==> 3 4 // SPE1 // 3 // rejects <==> mulls response

Finally, consider the example in sentence n.670 again from Headlines,

//No plan to shut petrol pumps at night Moily.
//India govt rejects proposal to shut petrol pumps at night
1 2 <==> 3 4 // SIMI // 4 // No plan <==> rejects proposal

We set up a specialized algorithm for cases like the one in 214. But the other two cases we found are not computable: “no plan” can be paraphrased in an infinite number of different ways. The same applies to “mulls response”.

Coming now to the need to compute predicate-argument structures, we have been convinced of its usefulness only after discovering poor performance of the system in total NOALI classification. In all those cases in which the sentence pair did not share any chunk, the system was still trying to relate far-fetched similarities, despite the fact that the overall meaning was not compatible with that interpretation. Take for instance the pair from the test-set, sentence no.9:

// Many dead as asylum boat sinks off Australia
// Mandela spends third day in hospital
3 4 5 <==> 5 6 // SIMI // 2 // as asylum boat <==> in hospital

Our system wrongly found some similarity between two chunks, where “asylum” and “hospital” share meaning components. However, the two sentences are clearly talking about totally different topics so the apparent meaning similarity should not apply. Preventing this from happening could only take place in case predicate argument structures were made available. Main predicates SINK and SPEND would then be judged not to share any meaning nor would the SUBJect “asylum” and “Mandela”. Argumenthood could then be used to prevent a SUBJect “asylum boat” from being made to share semantic similarity components with a locative adjunct “in hospital”, where the sinking event doesn’t find any correspondence.

2.1 Matching procedures

Matching is applied at different levels using different resources. We use WordNet[2] for EQUI relations by searching non identical predicates in the same synset. Again WordNet is used for SIMI by searching up and down the path one level only, for all non identical predicates; we also use an extended version of VerbOcean[3] for entailment relations between verbs, where we added some 250 new relations. For more general REL relations we use Roget's Thesaurus[4]. And of course there's a great number of gazeteers and lexica for NER that is made available, in particular, the ones by JRC made available by Ralph Steinberg[5].

For more complex semantic similarity relations, we have reconstructed our Finite State Automaton that we introduced in our last participation in RTE n.5. We report it here below. It requires tagged words and a number of linguistic rules to be implemented. The current version of the algorithm is made up of 86 different rules.

The procedure takes as input the tagged list of words making up the current chunk pairs and tries to match it, by the predicate match_template(Chunk1, Chunk2). If the match succeeds the semantic evaluation outputs a value that is indicative of the type of decision taken. This matching procedure is reached by the analysis only after EQUI have failed. Consider the example below where we highlight the portion of the chunk pair relevant for the semantic evaluation:

T: Trains, trams, cars and buses ground to a halt on Monday after a shoot-out between 18:00 CET and 19:00 CET in the historical *city of Basel in Switzerland*.
H: *Basel is a European city*.

In more detail, Augmented Finite State Automata mean that in addition to equality matching that is at the basis of the whole algorithm, the system looks for inferences and other lexical information to authorize the match. In fact, these procedures as a whole allow the matching to become more general though introducing some constraints. The instructions reported below are expressed in Prolog which treats capitalized words as variables. Constants on the contrary are written with lower case letter, as for instance the words "of" and "in" below.

```
match_template([A,Is-_,T-_,F_,G|Hyp],
[G,of-_,A,in-_,L-_|Text])
:-
lightverbs(Is),
high_rank(T,Lex),
locwn(L),
is_in(L,F1),
(natl(F1,F,_);natl(F1,_,F)), !.
```

where the procedure "lightverbs" looks for copulative verbs, i.e. the verb of the Chunk1 must be a copulative verb; "high_rank" looks for high frequency words like articles; "locwn" verifies that the word present in the variable "L" is a location. Then there are two inferences: the first one is fired by the call "is_in" that recovers the name of the continent to which "L" belongs, thus implicitly requiring "L" to be the name of a nation. Then the second inference looks for the corresponding nationality adjective. Values for all above variables (L,A,F,F1,T,G,Is), are then as follows:

L --> Switzerland, F1 --> Europe, F --> European
A = Basel-np, Is = is, T = a, G = city-n

3 Analysing iSTS Annotation Criteria

As said above, measuring semantic similarity between two sentences (snippets) resembles very closely TE. However, differences are clearer seen that iSTS imposes an additional first step - chunk alignment - which is finding the chunk pair that is semantically closer and then assigning a type and a score. In this sense, it is intuitively limited to what pertains to semantics, while TE had no such limitation and the word Entailment was understood as possessing a much wider import than just semantics. In fact, if we read the annotation guidelines carefully, we discover at pag. 2 that the subdivision into chunks is defined as follows: "Chunks are aligned in context, taking into account the interpretation of the whole sentence, including common sense." Common sense has no reference whatsoever to SEMANTICS being rather based on knowledge of the world. The same we find at pag. 3, "Note that the interpretation of the whole sentence, including common sense inference, has

[2] https://wordnet.princeton.edu/wordnet/download/current-version/
[3] http://demo.patrickpantel.com/demos/verbocean/
[4] http://www.gutenberg.org/ebooks/10681
[5] http://langtech.jrc.it/RS.html

to be taken into account." Then in the annotation guidelines, we find other elements that indicate a need to go beyond semantics: "B. When aligning, take into account the deep meaning of the chunk in context, beyond the surface."

In fact, we expected task data to pertain to so-called "literal" and direct meaning decomposition subset of sentences and not to contain any "non-literal" or "indirect" interpretable data. We also didn't expect to find unexpressed or implicit meaning components that had to be reconstructed while interpreting chunks. As we will see, this is only partially confirmed. The task itself has been organized so as to favour semantic decomposition operations to be applicable to chunks which should be at first paired by the system and then interpreted and "semantically explained". Semantic explanation is to be carried out by choosing among a small number of semantic relational label, including:

- EQUI(valent) SIMI(lar) OPPO(site) SPE(cific)1 SPE(cific)2 REL(ation) NOALI(gnment)

but also additional labels with further semantic content:

- EQUI_FACT EQUI_POL SIMI_FACT SPE1_FACT SPE2_FACT NOALI_FACT SIMI-2

In particular label SIMI-2 is not explained in the guidelines. As to the other additional labels, FACT stands for factuality and POL for polarity, i.e. these two extensions should be used in case the sentence contains elements of one or the other phenomenon.

The 2016 version of the task introduces then some further difficulty, when it allows chunks to match not just in a 1:1 but also 2:1. As to this procedure, in our system we check for a possible match already in the first recursive search and at the end of the recursive matching procedure.

We will now look into semantic relation labels first and see whether they are adequate and consistent. The first label, EQUIvalent covers all cases of full identity between wordforms in the two chunks. Equivalence without full orthographic identity is very frequent and includes all cases where the system matching algorithm has to put up with capitalized, or fully upper case words, and sometimes dashed version of the same unique word. But certainly the most frequent cases of equivalence-not-identity are where the two wordforms have different morphology and lemmata have to be matched. During EQUI matching procedures, we transform head words into their corresponding lemmata and try a match.

More difficult cases include named entity recognition processes, whenever an institution or a person is present with the abbreviated wordform in one chunk and the fully expressed name in the other; or when the name is used in one chunk and the other contains name and surname. Eventually, in some cases, the nationality has to match the word for the nation in the other chunk. Differentiating between SIMI and SPE is certainly hard[6].

In all these definitions, we find the same words "similar meanings" except for REL where the word "relation" is used instead. Now differences between SPE and SIMI can only be found in the presence of "attributes" in the definition of SIMI, whereas in SPE we find reference to specificity, which we may assume can be related to the presence of "attributes" but in different measure. It would seem then that SPE relations are more "similar" than SIMI relations, where semantic relations are less close. Let's now look at the examples presented in the guidelines. The second one poses already a problem:

[Red double decker bus][driving][through the streets]
[Double decker passenger bus][driving][with traffic]
Alignment: 1<==>1(SPE1 4), 2<==>2(EQUI 5), 3<==>3(REL 3)

The first two chunks have been interpreted as entertaining a SPE1 relation, which means that the first chunk "Red double decker bus" is more SPEcific than the second one. But the second one also contains a specification which is not present in the first chunk, "double decker passenger bus", constituted by the noun modifier "passenger". The choice of regarding chunk 1 more specific is driven by the fact that the colour specification adds a more relevant information to the common "double decker bus" than the noun "passenger", which is regarded of no import to the identification of the semantic reference realized by the multiword head double_decker_bus. So it would seem that in order to decide whether to use SPE1 or SPE2 a system for semantic evaluation should be equipped with "commonsense" knowledge that would weigh the two attributes accordingly. However, it may be

[6] The guidelines defines it as follows (pag.3/4) very vaguely.

disputable to define reference to "colour" as a less relevant attribute than reference to "passenger", for the simple reason that this decision eventually depends on the spatial location of the event. If it is England the location, then it is a commonplace notion that double decker buses are just red. But suppose the location was Lisbon, where double decker are sometimes red sometimes multicoloured, these latter being used by tourists to tour the city. In that case colour would have been more relevant than passenger. Even though "passenger" constitutes a more general attribute than "tourist" bus. So eventually, in order to use commonsense knowledge, at least time/place location must be made available.

Similar problems may be raised in another example (pag.6), where the semantic relation is expressed by predicates:

[Hundreds]1[of Bangladesh clothes factory workers]2 [ill]3
[Hundreds]1[fall]2[sick]3[in Bangladesh factory]4
Alignment: 1<==>1 (EQUI 5), 2<==>4 (SPE1 3), 3 <==> 2,3 (EQUI 5)

Here the adjective "ill" is make to relate to "fall sick" by EQUI. However, (TO BE) "ill" where the verb to be is simply left implicit in the nominalized title, is interpretable as a STATE; whereas TO "fall sick" is clearly an EVENT. So maybe the two predicates are not EQUI but SIMI and in order to align FALL to the missing BE some inference is required. This goes against what is being affirmed under 2b, as to the fact that two events are (weakly) relatable but cannot be aligned because they "refer to different events". However the reference to different events is not clearly inferrable.

[Saudis]1 [to permit]2 [women]3 [to compete]4 [in Olympics]5
[Women]1 [are confronting]2 [a glass ceiling]3
Alignment: 1<==>Ø(NOALI), 2<==>Ø(NOALI), 3<==>1(SPE1 4), 4<==>Ø(NOALI), 5<==>Ø(NOALI), Ø<==>2(NOALI), Ø<==>3 (NOALI)

In one sentence we are told that women are in Saudi Arabia, but nothing is said in the second sentence, and since spatio-temporal locations can be left implicit we are unable to separate the two events and the two references to women. So we find it hard to consider the semantic relation intervening between the two chunks as SIMI.

4 Results

We report here below results for two of the three corpora only. As to the corpus for student-answer, it was filled with typos and spelling errors, and this was regarded part of the task. We were unable to compute any reasonable semantic similarity match for an extended number of sentences. So we decided to abandon it. As to the other two texts, results for test texts are very similar to those we already obtained for training ones, so we only show test results.

	Ali	Type	Score	TypSco
Venseval	0.7428	0.4667	0.6949	0.4624
Baseline	0.7100	0.4043	0.6251	0.4043

Table 1: Results compared to Baseline for IMAGES SYS – Rank 12 over 13

	Ali	Type	Score	TypSco
Venseval	0.8443	0.5789	0.8046	0.5735
Baseline	0.8556	0.4799	0.7456	0.4799

Table 2: Results compared to Baseline for IMAGES GS – Rank 10 over 20

As can be noticed by comparing results in SYS and GS tables, in the analysis of Images corpus the decrease of performance of the system is higher than the difference between baselines. Here below results for Headlines.

	Ali	Type	Score	TypSco
Venseval	0.7081	0.4679	0.6493	0.4531
Baseline	0.6486	0.4379	0.5912	0.4379

Table 3: Results compared to Baseline for HEADLINES SYS – Rank 12 over 13

	Ali	Type	Score	TypSco
Venseval	0.8731	0.5927	0.8099	0.5729
Baseline	0.8462	0.5462	0.7610	0.5461

Table 4: Results compared to Baseline for HEADLINES GS – Rank 13 over 20

In the Headlines corpus analysis differences in performance between SYS and GS are comparable. If we look closer at results in terms of number of

teams we see that they are only 8, and our rank is now fifth, both in Images and Headlines results. In conclusion, we favoured a rule-based approach because we assume it can account for differences in text structures. However, rules require fine-tuning which cannot be completed in a short time. This is clearly born out by differences in performance between the two corpora analysed, Images and Headlines, where the first one should have been much easier to process than the second.

References

Agirre, E. and Banea, C. and Cardie, C. and Cer, D. and Diab, M. and Gonzalez-Agirre, A. and Guo, W. and Lopez-Gazpio, I. and Maritxalar, M. and Mihalcea, R. and Rigau, G. and Uria, L. and Wiebe, J. (2015). SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), June.

Chklovski, Timothy & Patrick Pantel. 2004. VERBOCEAN: Mining the Web for Fine-Grained Semantic Verb Relations. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04). Barcelona, Spain.

Delmonte R., S.Tonelli, R. Tripodi, 2009. Semantic Processing for Text Entailment with VENSES, in Proceedings of Text Analysis Conference (TAC) 2009 Workshop - Notebook Papers and Results, NIST, Gaithersburg MA, pp. 453-460.

Fellbaum, Christiane (ed.), 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Levy, Omer, Torsten Zesch, Ido Dagan, and Iryna Gurevych. 2013. Recognizing Partial Textual Entailment. In *ACL (2)*, pages 451–455.