SEMEVAL-2015-Task2 -Interpretable STS Annotation Guidelines

Introduction

The present guidelines have been written for the Semeval 2015 pilot subtask on Interpretable Semantic Textual Similarity. The pilot explores whether participant systems are able to explain **WHY** they think the two sentences are related / unrelated, adding an explanatory layer to the similarity score. As a first step in this direction, given a pair of sentences, participating systems will need to **align the chunks** in s1 to the chunks in s2, describing what **kind of relation** exists between each pair of chunks, and a **score** for the similarity/relatedness between the pair of chunks.

Chunks are aligned in context, taking into account the interpretation of the whole sentence, including common sense.

This report is organized as follows. We first define the chunks. We then give the main annotation steps, followed by the general guidelines, the instructions for scores and the instructions for assigning labels. We then provide specific guidelines with detailed information and examples. Finally, we include two sections about the actual implementation, including the interface and the detailed annotation procedure.

Definition of chunks

We follow Abney (1991) to define the chunks which will be aligned. According to Abney (1991), a chunk is "*a non-recursive core of an intra-clausal constituent, extending from its beginning to its head. A typical chunk consists of a content word surrounded by a constellation of function words, matching a fixed template*".

> [I begin] [with an intuition]: [when I read] [a sentence], [I read it] [a chunk] [at a time].
> [The bald man] [was sitting] [on his chair]

We follow the CONLL 2000 guidelines[1], adapting them to our purpose: 1) We split the main clause in smaller chunks (NPs, verb chains, PPs, adverbs and expressions), while simple subordinate clauses are kept as a whole chunk. 2) We take PPs as whole chunks [of that man].

Here you are some examples of our chunking:
- **NP** [The girl] / [Bradley Cooper and JJ Abrams]
- **verb chain** [is arriving] / [does not like]
- **PP** [at a time] / [with the telescope] / [the house] [of that man]
- **adverbs** [of course]
- **subordinate clauses** [When I read] [I] [am] [happy]
- **expressions** [once upon a time] / [by the way]

In order to help the annotators, we run the sentences through a chunker[2] trained on CONLL 2000 corpora.

Annotation steps

The main steps are as follows:
1. First identify the chunks in each sentence separately (in paper), regardless of the corresponding sentence in the pair.
2. Align chunks in order, using the interface (see Interface Section below), from the clearest and strongest correspondences to the most unclear or weakest ones.
3. For each alignment, provide a similarity/relatedness score (see Similarity and Relatedness score Section below).
4. For each alignment, choose one (or more) alignment label (see Labels for alignment Section below).

The detailed procedure is specified in the Procedure Section below.

General guidelines

A. Each sentence pair is independent of the other sentences in the dataset.
B. When aligning, take into account the deep meaning of the chunk in context, beyond the surface.
C. Try to align as many chunks as possible.
D. In the current pilot, one chunk can be aligned with at most one chunk (1:1 relation), i.e. it cannot be aligned to two chunks.
E. When having two options to align, choose the strongest corresponding chunk first. As a workaround, the other chunk will be left unaligned, marked with a special label (ALIC, see below).
F. Chunks can be left unaligned if no corresponding chunk can be found, or if the corresponding chunk has been already aligned (NOALI and ALIC labels respectively, see below).
G. Assign at least one label to each alignment.
H. You can leave punctuations unaligned, as they will be ignored when evaluating. The interface requires that you annotate all tokens, so please tag them with the label for unaligned chunks.

---

[1] http://www.clips.ua.ac.be/conll2000/chunking/
[2] https://github.com/ixa-ehu/ixa-pipe-chunk

Similarity and Relatedness score

Independently of the labels, and **before** assigning **any label, please provide a similarity/relatedness score** for each alignment from 5 (maximum similarity/relatedness) to 0 (no relation at all), as follows:

- 5 if the meaning of both chunks is equivalent
- [4,3] iff the meaning of both chunks is very similar or closely related
- [2,1] iff the meaning of both chunks is slightly similar or somehow related
- 0 iff the meaning of both chunks is completely unrelated.

Note that you would never have a 0 for an aligned pair, as that would mean that the two chunks would be left unaligned. Note also that if the score is 5, then the label assigned later should be EQUI (see below). After assigning the label the annotator should check for the following:

- NOALI and ALIC should not have scores.
- EQUI should have a 5 score.
- The rest should have a score bigger than 0 but lower than 5.

Labels for alignment

The general labels for alignment are the following ones. Note that the **interpretation of the whole sentence, including common sense inference, has to be taken into account**. This means that we need to take into account the context in order to know whether they refer to the same instance (or set of instances) or not. Instances may refer to physical or abstract object instances (for NPs) or real world event instances (for verb chains):

1. **EQUI**: both chunks have the same meaning, they are semantically equivalent in this context.
2. **OPPO**: the meanings of the chunks are in opposition to each other, lying in an inherently incompatible binary relationship.
3. **SPE1**: both chunks have similar meanings, but chunk in sentence 1 is more specific or gives more details.
4. **SPE2**: like SPE1, but it is the chunk in sentence 2 which is more specific.

In addition, the meaning of the chunks can be very close, either because they have a similar meaning, or because their meanings have some other relation. In those cases, we use SIMI or REL as follows:

5. **SIMI**: both chunks have similar meanings, they share similar attributes and there is no EQUI, OPPO, SPE1, or SPE2 relation.
6. **REL**: both chunks are closely related, by some relation not mentioned above (i.e. no EQUI, OPPO, SPE1, SPE2, or SIMI relation).

Finally, the chunk can be left unaligned, as follows:

7. **ALIC**: this chunk has not any corresponding chunk in the other sentence because of the restriction on having one-to-one alignments[3], but otherwise the chunk would be aligned to some other chunk.
8. **NOALI**: this chunk has no corresponding chunk in the other sentence.

The **above eight labels are** exclusive, and each alignment should have one and only one such label. Note that the two last labels apply for chunks which are left unaligned.

In addition to one of the labels above, there are two labels which can be used in isolation or together, that is, you can use none, one or both:

A. **FACT**: the factuality in the aligned chunks (i.e. whether the statement is or is not a fact or a speculation) is different.
B. **POL**: the polarity in the aligned chunks (i.e. the expressed opinion, which can be positive, negative, or neutral) is different.

Note that ALIC and NOALI can also be FACT or POL, meaning that the respective chunk adds a factuality or polarity nuance to the sentence.

Specific guidelines with examples

In this Section we detail the guidelines providing some illustrative examples. For easier illustration, the alignments in the examples are shown in left-to-right order. However, the annotator follows a different order, as (s)he annotates the strongest alignments first. For instance, in the following example, the temporal order of annotations was the following:

$[12]_1$ $[killed]_2$ $[in\ bus\ accident]_3$ $[in\ Pakistan]_4$
$[10]_1$ $[killed]_2$ $[in\ road\ accident]_3$ $[in\ NW\ Pakistan]_4$

Order in which the human annotator decides the alignments:
2 ⇔ 2 (EQUI 5),
3 ⇔ 3 (SPE1 4),
4 ⇔ 4 (SPE2 4),
1 ⇔ 1 (SIMI 4)

Order in which we report the annotations in this section:
Alignment of chunks: 1 ⇔ 1 (SIMI 4), 2 ⇔ 2 (EQUI 5), 3 ⇔ 3 (SPE1 4), 4 ⇔ 4 (SPE2 4)

For each alignment we specify the similarity/relatedness score in red.

---

[3] see *D in General guidelines*

1.- Align chunks which have the same or related meaning taking into account the context and interpretation of the corresponding sentence. The examples below include one sample alignment for each possible alignment label:

[Red double decker bus]$_1$ [driving]$_2$ [through the streets]$_3$
[Double decker passenger bus]$_1$ [driving]$_2$ [with traffic]$_3$
Alignment of chunks: 1 ⇔ 1 (SPE1 4), 2 ⇔ 2 (EQUI 5), 3 ⇔ 3 (REL 3)

[2 car bombs]$_1$ [kill]$_2$ [8]$_3$ [in southern Iraq]$_4$
[Car bombing]$_1$ [kills]$_2$ [14]$_3$ [in northern Iraq]$_4$
Alignment of chunks: 1 ⇔ 1 (SPE1 4), 2 ⇔ 2 (EQUI 5), 3 ⇔ 3 (SIMI 3), 4 ⇔ 4 (OPPO 4)

[Stocks]$_1$ [soar]$_2$ [on Wall St lead]$_3$
[Stocks]$_1$ [slump]$_2$ [on Wall Street]$_3$
Alignment of chunks: 1 ⇔ 1 (EQUI 5), 2 ⇔ 2 (OPPO 4), 3 ⇔ 3 (SPE1 3)

2.- In some cases, it is necessary to understand the events described in the sentences and the roles played by the chunks to be aligned. Usually, the aligned chunks play similar roles in the underlying event:
[Mall attackers]$_1$ [used]$_2$ [`less is more' strategy]$_3$
[In Kenya]$_1$, [attackers]$_2$ [used]$_3$ [`less is more' strategy]$_4$
Alignment of chunks: 1 ⇔ 2 (SPE1 4), 2 ⇔ 3 (EQUI 5), 3 ⇔ 4 (EQUI 5), Ø ⇔ 1 (NOALI)

[Gunmen]$_1$ [abduct]$_2$ [seven foreign workers]$_3$ [in Nigeria]$_4$
[Seven foreign workers]$_1$ [kidnapped]$_2$ [in Nigeria]$_3$
Alignment of chunks:1 ⇔ Ø (NOALI), 2 ⇔ 2 (EQUI 5), 3 ⇔ 1 (EQUI 5), 4 ⇔ 3 (EQUI 5)

[A very clear miscarriage of justice]$_1$
[I]$_1$ [agree on]$_2$ [the miscarriage of justice]$_3$
Alignment of chunks: 1 ⇔ 3 (SPE1 4), Ø ⇔ 1 (NOALI), Ø ⇔ 2 (NOALI)

2a. When the chunks play different but related roles also align them:

[Hundreds]$_1$ [**of Bangladesh clothes factory workers**]$_2$ [ill]$_3$
[Hundreds]$_1$ [fall]$_2$ [sick]$_3$ [**in Bangladesh factory**]$_4$
Alignment of chunks: 1 ⇔ 1 (EQUI 5), **2 ⇔ 4 (SPE1 3),** 3 ⇔ 3 (EQUI 5), Ø ⇔ 2 (ALIC)

[Man]$_1$ [**in yellow canoe**]$_2$ [paddling]$_3$ [through water]$_4$
[Man]$_1$ [paddling]$_2$ [**a yellow canoe**]$_3$ [towards a the shore]$_4$
Alignment of chunks: 1 ⇔ 1 (EQUI 5), **2 ⇔ 3 (EQUI 5)**, 3 ⇔ 2 (EQUI 5), 4 ⇔ 4 (SPE2 3)

2b.- **When** the sentences **refer to different events**, then the roles can be different, e.g. you should align **3 ⇔ 1** in the next sentence pair:

[Saudis]₁ [to permit]₂ [**women**]₃ [to compete]₄ [in Olympics]₅
[**Women**]₁ [are confronting]₂ [a glass ceiling]₃
Alignment of chunks: 1 ⇔ Ø (NOALI), 2 ⇔ Ø (NOALI), **3 ⇔ 1 (EQUI 5)**, 4 ⇔ Ø (NOALI),
        5 ⇔ Ø (NOALI), Ø ⇔ 2 (NOALI), Ø ⇔ 3 (NOALI)

3.- Sometimes the one-to-one restriction will leave some chunks with no alignment, even if there is a clear correspondence to one of the already aligned chunks, or even if It is clear that the meaning of the chunk can be inferred from the other sentence. In these cases, use ALIC (aligned to context):

[**Saudis**]₁ [to permit]₂ [women]₃ [to compete]₄ [in Olympics]₅
[Saudi Women]₁ [Allowed]₂ [To Compete]₃ [At Olympics]₄
Alignment of chunks: **1 ⇔ Ø (ALIC)**, 2 ⇔ 2 (EQUI 5), 3 ⇔ 1 (EQUI 5), 4 ⇔ 3 (EQUI 5), 5 ⇔ 4 (EQUI 5)

In contrast to the above, if the pair was the following, then we would have 1 ⇔ 1 (SPE2 3):
        [**Saudis**]₁ [to compete]₂ [in Olympics]₃
        [**Saudi Women**]₁ [Allowed]₂ [To Compete]₃ [At Olympics]₄

Another example:

[Islamists]₁ [attack]₂ [north **Mali** city]₃ [after suicide bombings]₄
[Islamists]₁ [attack]₂ [**Malian troops**]₃ [in Gao]₄
Alignment of chunks: 1 ⇔ 1 (EQUI 5), 2 ⇔ 2 (EQUI 5), 3 ⇔ 4 (SPE2 4), 4 ⇔ Ø (NOALI),
        Ø ⇔ 3 (ALIC)

In contrast to the above, if the pair was the following, then we would have 3 ⇔ 3 (REL 2):
        [Islamists]₁ [attack]₂ [north **Mali** city]₃
        [Islamists]₁ [attack]₂ [**Malian** troops]₃

Another example:

[2 dead]₁, [2 injured]₂ [in **Nevada** middle school shooting]₃
[**Nevada**]₁: [2 dead]₂, [2 hurt]₃ [in middle school shooting]₄
Alignment of chunks: 1 ⇔ 2 (EQUI 5), 2 ⇔ 3 (EQUI 5), 3 ⇔ 4 (EQUI 5), Ø ⇔ 1 (ALIC)

In contrast to the above, if the second sentence was the following, then we would have 3 ⇔ 1 (SPE1 2):
        [**Nevada**]₁: [2 dead]₂, [2 hurt]₃

4.- Spelling errors will be ignored when they do not affect the meaning of the sentence, and they will be therefore annotated as if there were no errors:

[People]$_1$ [sitting]$_2$ [on the porch]$_3$
[People]$_1$ [sitting]$_2$ **[on acouch]**$_3$
Alignment of chunks: 1 ⇔ 1 (EQUI 5), 2 ⇔ 2 (EQUI 5), 3 ⇔ 3 (REL 2)

[sheep]$_1$ [standing]$_2$ **[in afield]**$_3$
[A sheep]$_1$ [grazing]$_2$ [in a field]$_3$
Alignment of chunks: 1 ⇔ 1 (EQUI 5), 2 ⇔ 2 (SPE2 3), 3 ⇔ 3 (EQUI 5)

Interface

We reused the LDC word alignment interface[4], originally designed for machine translation. We added several buttons to comply with the labels, and added an extra slot for the similarity/relatedness score.

The interface.

The annotation labels in the interface.

Procedure

The annotator will proceed step by step as follows:

1. Using the automatically chunked version of the sentences (*.chunk2 files), identify the chunks in each sentence separately, and write them in paper. Note that you should not think on alignments yet (in fact, you should not even read the other sentence).
2. Identify the alignments in paper.
3. Go to the interface
   a. choose files with 1st and 2nd sentence (*.sent1.txt, *.sent2.txt)
   b. create output gold standard file (*.wa) (optionally you can open a previously created output gold standard file *.wa)
   c. for each sentence pair, proceed from strongest to weakest alignments:
      i. Tick on the tokens of each chunk
      ii. Type in the similarity/relatedness number
      iii. Choose the alignment labels:
         1. main label (among EQUI, OPPO, SPE1, SPE2, SIMI, REL, ALIC, NOALI)
         2. optionally choose FACT
         3. optionally choose POL
      iv. Check that all tokens have been used
   d. Go to c
4. Double check that the chunks derived from the gold standard (*.chunk2.gs) match those in paper.
5. Double check that all OPPO, SPE1, SPE2, SIMI, REL, alignments have a score (non-5, non-0 score), and EQUI has a 5 score.

References

Abney, S. (1991). Parding By Chunks. In: Robert Berwick, Steven Abney and Carol Tenny (eds.), Principle-Based Parsing. Kluwer Academic Publishers, Dordrecht.

Authors

Eneko Agirre
Montse Maritxalar
German Rigau
Larraitz Uria