SEMEVAL-2016, Interpretable STS Annotation Guidelines

V 2.2 October 28, 2015

Authors: Eneko Agirre, Montse Maritxalar, German Rigau, Larraitz Uria

Updates with respect to SemEval-2015

(To be read by those familiar with SemEval-2015 guidelines)

In the pilot subtask presented in SemEval-2015, the alignment of chunks was restricted to the 1:1 relation, that is, a chunk could be aligned with at most one chunk, it could not be aligned to two chunks. When there were two options to align, the strongest corresponding chunk was first chosen and the other chunk was left unaligned, marked with a special label: ALIC. Therefore, in some cases a chunk had not any corresponding chunk in the other sentence because of the restriction on having one-to-one alignments, but otherwise the chunk would have been aligned to some other chunk. In the current task, ALIC has disappeared and the chunk that would have been left unaligned last year is now aligned. To align those chunks, see F in the *General guidelines* section above, and the *Specific guidelines* section.

We have also decided to split the subordinate clauses in smaller chunks.

Regarding the corpus we have added a new domain (see *specific guidelines for student answers corpus* section).

On top of that, we refined and revised some of the explanations.

Introduction

The present guidelines have been written for the Semeval-2016 task on Interpretable Semantic Textual Similarity. The task explores whether participant systems are able to explain **WHY** they think 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 sentence1 to the chunks in sentence2, adding a **score** for the similarity/relatedness between each pair of chunks and describing what **kind of relation** exists between them.

Chunks are aligned in context, taking into account the interpretation of the whole sentence, including common sense. Our goal is to find chunk-level alignments whenever possible, and label those alignments. We do not aim at aligning and labelling longer phrases.

This report is organized as follows. We first define the chunks. We then give the main annotation steps, the instructions for scores and the instructions for assigning labels, and the general guidelines. Then, we 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

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*". In 1996, Abney reformulated his definition in terms of *islands of certainty*, providing a more flexible definition which is applicable to more languages: '*a chunk is an intra-clausal constituent including pre-head as well as post-head modifiers, but not pp-attachment or sentential elements'*.

[The bald man] [was sitting] [on his chair]

We take into account Abney (1996) to define the chunks and we also follow the CONLL 2000 guidelines[1], adapting them to our purpose:

1) We split the main clause and subordinate clauses in smaller chunks (NPs, verb chains, PPs, adverbs and expressions),

2) We take PPs as whole chunks.

Here you have 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]
- **expressions** [once upon a time] / [by the way]

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

[1] http://www.clips.ua.ac.be/conll2000/chunking/

[2] https://github.com/ixa-ehu/ixa-pipe-chunk

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, but we will first present the scores, labels, general guidelines and specific guidelines.

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 (represented as NIL) if the meaning of the chunk 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 should have NIL score.
- EQUI should have a 5 score.
- The rest of the labels 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 the **aligned chunks 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.
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 not considered similar but they are closely related by some relation not mentioned above (i.e. no EQUI, OPPO, SPE1, SPE2, or SIMI relation).
7. **NOALI**: this chunk has not any corresponding chunk in the other sentence. Therefore, it is left unaligned.

The **above seven labels are exclusive**, and each alignment should have one such label.

In addition to one of the labels above, there are two labels which can be used either 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 NOALI can also be FACT or POL, meaning that the respective chunk adds a factuality or polarity nuance to the sentence.

General guidelines

These are the general guidelines, which give a general idea of the process. These are underspecified and are given as a short introduction. Please read the specific guidelines for guidance.

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. One chunk can be aligned to more than one chunk, but only to prevent unaligned chunks.
D. Do all 1:1 alignments first. When having two options to align, choose the strongest corresponding one first.
E. After doing 1:1 alignments, check unaligned chunks. There are three options to align them, in this order of preference:
    1. Insert the unaligned chunk (or group of chunks) into an existing 1:1 alignment.
    2. Create a new relation, add a new score and label to the new relation.
    3. Chunks can be left unaligned if no corresponding chunk can be found.
F. Assign at least one label to each alignment.
G. Try to leave as few unaligned chunks as possible.
H. Keep it simple.
I. 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.

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.

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:

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

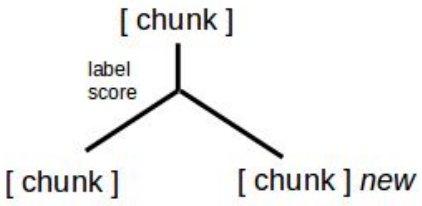
[Hundreds]$_1$ [**of Bangladesh clothes factory workers**]$_2$ [ill]$_3$
[Hundreds]$_1$ [fall]$_2$ [sick]$_3$ [**in Bangladesh factory**]$_4$
Alignment of chunks: 1 ⇔ 1 (EQUI 5), **2 ⇔ 4 (SPE1 3),** 3 ⇔ 2,3 (EQUI 5)

2b.- When the sentences **refer to different events**, then the chunks can be aligned even if the roles are different. In this case, the aligned chunks need to be closely related (labels EQUI, SIMI, SPE). For instance, you should align **3 ⇔ 1** in the next sentence pair (where “women” in the first sentence refers to “saudi women” and is thus more specific), but not the rest, even if there are some weak relations between them (e.g. “to compete” and “are confronting”):

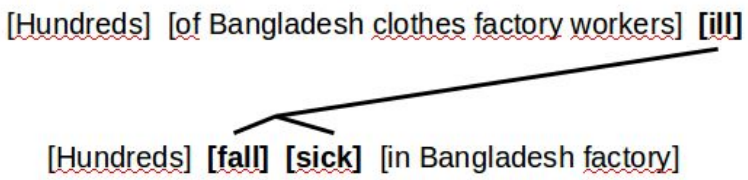
[Saudis]$_1$ [to permit]$_2$ [**women**]$_3$ [to compete]$_4$ [in Olympics]$_5$
[**Women**]$_1$ [are confronting]$_2$ [a glass ceiling]$_3$
Alignment of chunks: 1 ⇔ Ø (NOALI), 2 ⇔ Ø (NOALI), **3 ⇔ 1 (SPE1 4)**, 4 ⇔ Ø (NOALI),
5 ⇔ Ø (NOALI), Ø ⇔ 2 (NOALI), Ø ⇔ 3 (NOALI)

3.- As specified in D and E in the general guidelines above, after doing 1:1 alignments, check unaligned chunks. **There are three possibilities in this order of preference: fold in into an existing alignment, create a new alignment, or leave unaligned, as follows:**

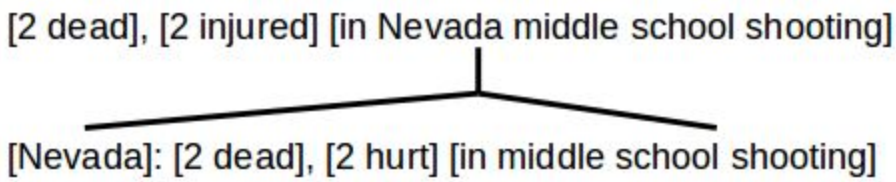
- 3a.- When the unaligned chunk is referred in one existing alignment, **the chunk can be folded in into that alignment**, but keeping the same score and label. If several options exist, always choose the strongest alignment first. Note that, given that the context needs to be taken into account, the chunk being incorporated was already considered when assigning the label and the score to the previous 1:1 alignment.



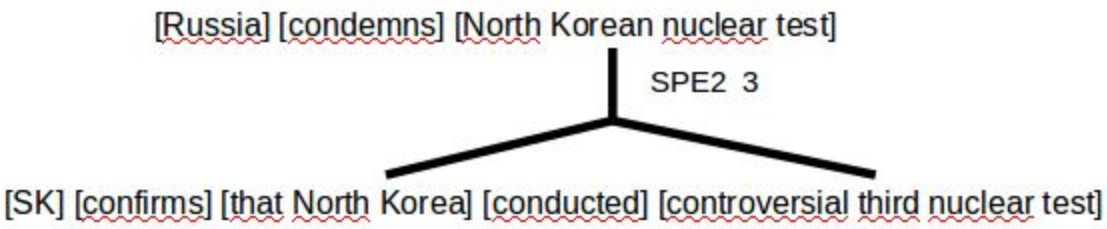
[Hundreds]$_1$ [of Bangladesh clothes factory workers]$_2$ **[ill]**$_3$
[Hundreds]$_1$ [**fall**]$_2$ [sick]$_3$ [in Bangladesh factory]$_4$
Alignment of chunks: 1 ⇔ 1 (EQUI 5), 2 ⇔ 4 (SPE1 3), **3 ⇔ 2,3 (EQUI 5)**
Note that, before folding in [fall]$_2$, 3 and 3 were aligned with EQUI 5.



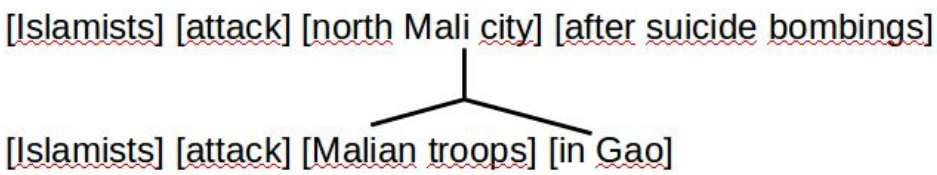
[2 dead]$_1$, [2 injured]$_2$ [in **Nevada** middle school shooting]$_3$
[**Nevada**]$_1$: [2 dead]$_2$, [2 hurt]$_3$ [in middle school shooting]$_4$
Alignment of chunks: 1 ⇔ 2 (EQUI 5), 2 ⇔ 3 (EQUI 5), **3 ⇔ 1,4 (EQUI 5)**
Note that, before folding in [Nevada]$_1$, 3 and 4 were aligned with EQUI 5.



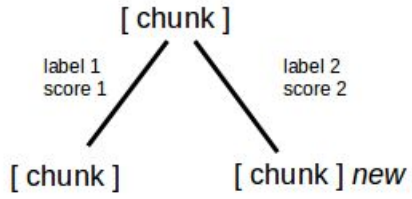
[Russia]$_1$ [condemns]$_2$ [**North Korean** nuclear test]$_3$
[South Korea]$_1$ [confirms]$_2$ [that **North Korea**]$_3$ [has conducted]$_4$ [controversial third nuclear test]$_5$
Alignment of chunks: 1 ⇔ 1 (SIMI 3), 2 ⇔ 2 (REL 4), **3 ⇔ 3,5 (SPE2 3)**, Ø ⇔ 4 (NOALI)
Note that, before folding in [that North Korea]$_3$, 3 and 5 were aligned with SPE2 3.



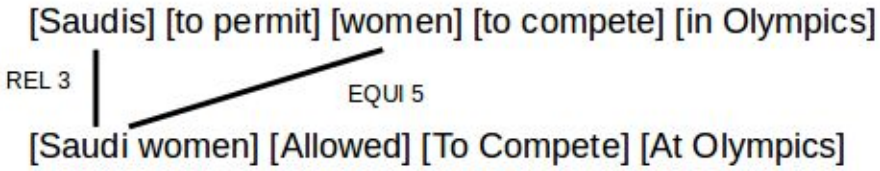
[Islamists]$_1$ [attack]$_2$ [north **Mali** city]$_3$ [after suicide bombings]$_4$
[Islamists]$_1$ [attack]$_2$ [**Malian troops**]$_3$ [in Gao]$_4$
Alignment of chunks: 1 ⇔ 1 (EQUI 5), 2 ⇔ 2 (EQUI 5), **3 ⇔ 3,4 (SPE2 3)**, 4 ⇔ Ø (NOALI)
Note that, before folding in [Malian troops]$_3$, 3 and 4 were aligned with SPE2 3.



- 3b.- In some exceptional cases, when the unaligned chunk is referred to in an existing alignment but plays a different role and it can't be folded in into the existing alignment, **it is necessary to create a new alignment**:



[**Saudis**]$_1$ [to permit]$_2$ [women]$_3$ [to compete]$_4$ [in Olympics]$_5$
[Saudi Women]$_1$ [Allowed]$_2$ [To Compete]$_3$ [At Olympics]$_4$
Align. of chunks: **1 ⇔ 1 (REL 3)**, 2 ⇔ 2 (EQUI 5), **3 ⇔ 1 (EQUI 5)**, 4 ⇔ 3 (EQUI 5), 5 ⇔ 4 (EQUI 5)
Note that we create a new alignment, 1 ⇔ 1 (REL 3), because [Saudis]$_1$ is playing a different role to [women]$_3$, and, therefore, it can not be folded in the existing 3 ⇔ 1 (EQUI 5) alignment.



- 3c.- When the unaligned chunk(s) can not be aligned, we leave the chunk(s) definitely unaligned (NOALI)

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)

Specific guidelines for the Student Answers corpus

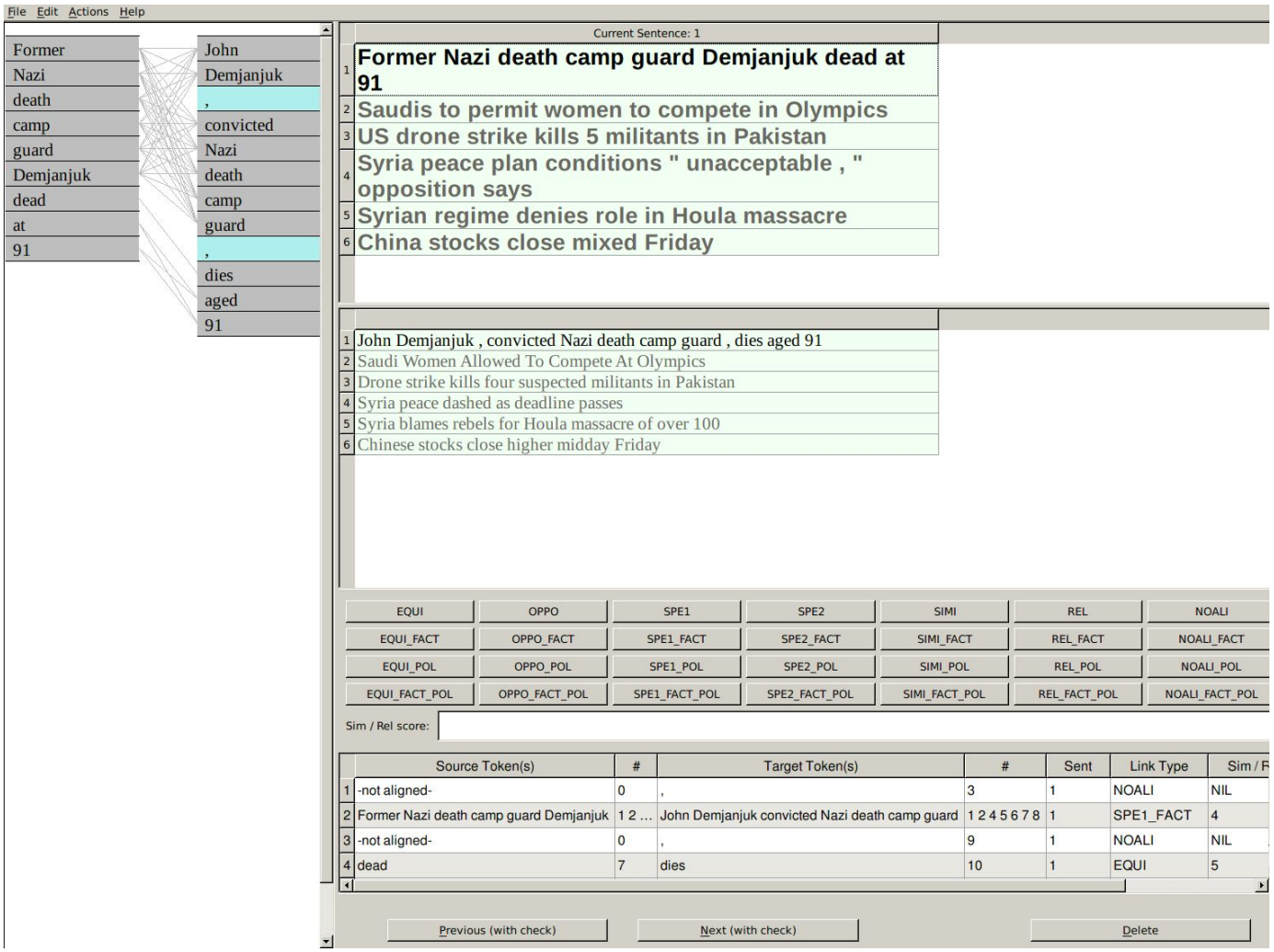
The student answers corpus consists of the interactions between students and the BEETLE II tutorial dialogue system (Dzikovska et al., 2012). The BEETLE II system is an intelligent tutoring engine that teaches students in basic electricity and electronics. Students answer to some questions about circuits. In the present corpus, we include sentence pairs composed of a student answer and the reference answer of a teacher. We have rejected those answers containing pronouns whose antecedent is not in the sentence (pronominal coreference), because, as the question is not included in our corpus, we can not deduce which is the antecedent.

There are some aspects which are specific to this corpus and have to be taken into account:

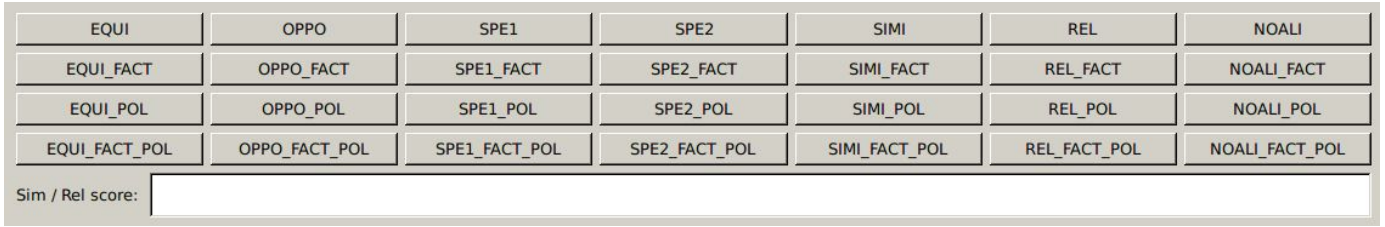
- A, B and C refer to bulb A, B and C.
- X, Y, and Z refer to switches X, Y, and Z.
- When numbers appear alone in a chunk, they refer to circuits.
- By default there is a unique battery, unless it is not explicitly mentioned.
- By default paths are considered to be closed.

Interface

We reused the LDC word alignment interface[3], 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.



[3] https://www.ldc.upenn.edu/language-resources/tools/ldc-word-aligner

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
    d. proceed from strongest to weakest 1:1 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)
            2. optionally choose FACT
            3. optionally choose POL
        iv. After finishing 1:1 alignments, check unaligned chunks and proceed as specified in point 3 in the specific guidelines: folding the unaligned chunk in an already existing alignment, creating a new alignment or leaving the chunk unaligned (main level NOALI, and optionally FACT and/or POL)
        v. Check that all tokens have been used
    e. 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.

Dzikovska et al (2012). Towards effective tutorial feedback for explanation questions: A dataset and baselines.