Semeval 2018-Task 7: Semantic Relation Extraction and Classification in Scientific Papers

Entity Annotation Guideline

Anne-Kathrin Schumann, Behrang QasemiZadeh
annek_schumann@gmx.de, zadeh@phil.uni-duesseldorf.de

Abstract

This document describes the annotation rules and process for entity annotation in the context of Semeval 2018-Task 7. The document is based on the ACL RD-TEC term annotation guidelines (Schumann and QasemiZadeh, 2015). In this document, annotators are assumed to possess basic knowledge of term annotation. Consequently, these guidelines explain only fine-grained annotation rules and the annotation process. For basic information on term annotation, please, refer to Schumann and QasemiZadeh (2015), section 1.

1 Annotation Rules

1.1 Semantic Classes

Terms can be sorted into different semantic classes. Table 1 gives an overview of semantic classes for computational linguistics terms. Annotators should use this table to decide whether a candidate is a term or not: *Only candidates that fall under one of the 7 categories should be annotated*. The classes are explained in more detail below:

- **Technology, System, and Method** terms refer to methods, processes, and approaches that are employed to solve practical tasks. In computational linguistics, "machine translation", "information extraction", "word sense disambiguation", ... are examples of technology terms. Often, the head noun is derived from a verb that describes a practical activity (e.g. "to analyze"

Semantic class	Examples
Technology, System, and Method	rule-based machine translation, transfer-based machine translation, interlingual machine translation, dictionary-based machine translation, RBMT, semantic analysis, speech synthesis, speech recognition, ...
Tool or Library	Stanford Core NLP, NLTK, OpenNLP library, Sphinx, CMUnicator, ...
Language resource	dictionary, grammar, transfer lexicon, syntactic rule, lexical selection rule, core grammar, ...
Language Resource Product	WordNet, Brown Corpus, SentiWordNet, Reuters-21578 , ...
Models	language model, translation models, word-based language model, error model, maximum entropy model, n-gram models, ...
Measures and Measurements	BLEU, NIST, Precision, Recall, F-Score, machine translation tests, MT evaluation, ...
Other Nominals (theories, formalism, ...)	target language, language, input sentence, output sentence, source language, orthographical variation, input, ambiguity, domain, lexical selection, Japanese, phrase, sentence, ...

Table 1: Semantic classes of terms with examples

and its nominalisation "semantic analysis") or process (e.g. "to propagate" and its nominalisation "(constraint) propagation"). In running text, such terms are sometimes accompanied by generic nouns such as "paradigm", "approach" or "method" (e.g. "rule-based machine translation paradigm").

- **Tool or library** terms refer to an actual implemented technology. Terms that belong to this class can be understood as instances of technology terms.[1] In the domain of computational linguistics, *tools* are often computer programmes used to carry out the actual analyses. For example, while "part of speech tagger" is a technology term, the terms "TreeTagger" and "Stanford PoS Tagger" belong to the category of *tools and libraries*.

- **Language resources** are mainly components of natural language processing (NLP) systems that contain linguistic knowledge, for example, lexical databases, corpora, and so on.

- **Language resource product** terms refer to actual language resources. For example, "Princeton WordNet" is a lexical database which can be obtained and used in a project.

[1] In other words, the *tool or library class* is a sub-category of the technology class.

- **Models** refer to method-specific knowledge resources. For example, "language model" is usually a database of the probability distribution of a sequence of words that is employed by a method to perform a task (e.g., part-of-speech tagging). Similarly, "phrase tables" and "translation models" are knowledge resources employed by particular types of machine translation technologies.

- **Measures and Measurements** comprise terms referring to measures (e.g. "BLEU", "f-score") or more abstract measurements (e.g. "accuracy" or "translation quality").

- Finally, the **other** are any other abstract concepts in the field of computational linguistics that cannot be fitted into any of the above listed classes. Therefore, this class of terms encompasses a large variety of terms such as linguistic entities, scientific disciplines and so on. Terms with a very specific meaning in a highly specialised context can also belong to the *other* category. For example, in the context of machine translation, *language pairs* such as "Czech-English" fall under the *other* class.

1.2 Determining Term Length

Annotators should make use of the following criteria when determining the length of a term:

- **Determiners**: *Do not* annotate determines or pronouns of any kind as part of a term. For example, in the string "a machine translation method", the term is "machine translation method".

- **Abbreviations**: Abbreviations can be terms if they designate specialized concepts, for example, RBMT or TF-IDF. These abbreviations are shorthands for expressing terms (e.g., RBMT for "rule-based machine translation").

- **Term-abbreviation sequence:** In the case that a term is followed by its abbreviation, the whole sequence is annotated as one term. For instance, given the text "machine translation (MT)", the whole sequence is annotated as one term (instead of annotating "machine translation" and "MT" as separate terms).

- **Terms broken by abbreviations**: In a number of occasions, an abbreviation of a general term is inserted into a term of more specific meaning. For example, in "machine translation (MT) evaluation", "MT" is the abbreviation of the term ""machine translation" which is inserted in the more

specific term "machine translation evaluation". In these circumstances, the whole sequence is annotated as one term. The semantic category of the term is decided by the semantics of the more specific term, namely "machine translation evaluation".

- **Proper nouns**: Proper nouns (names) should be annotated only if the related concept belongs to one of the categories listed in Table 1 such as the *Tool or Library* (e.g., TreeTagger, ABNER, ...), the *Language Resource Product* (e.g., EuroWordNet, WordNet, ...) or the *Other* categories (e.g., conferences such as LREC or associations such as ELRA, ACL, ...). Other kinds of names (e.g., people or place names) should not be annotated.

- **Generic nouns**: As explained before, terms can be accompanied by generic nouns, e.g., the word "approach" in "sequential labelling approach". In such a case, annotators are asked to include the generic noun into the term span (e.g., "sequential labelling approach"). We call this principle the **maximal length annotation** principle. For further help with generic nouns, also check sections 1.3 and 1.4.

- **Adjectival modifiers**: In many cases, terms are modified by adjectives, for example, in the strings "systematic pattern" and "statistical MT". How to decide whether the adjective is part of the term or not? Try by checking whether removing the adjective changes the meaning of the term or not. In the case of our examples, we see that saying "pattern" instead of "systematic pattern" does not change much since patterns are by default something that is systematic. "MT", however, is more general than "statistical MT" since there are also many other approaches to machine translation. In the first case, annotate the term *without* the redundant modifier (that is, "pattern" instead of "systematic pattern"). In the latter case, annotate the complete span of the noun phrase ("statistical MT" instead of "MT").

- **Conjunctions and prepositions**: Complex term phrases can contain conjunctions and prepositions, for example, "TREC 2003 and TREC 2004 QA tracks", or "automatic evaluation of machine translation and document summarization". In these cases, a set of rules applies:

  - For conjunctions, if the noun phrases linked by them are *ellipses*, the whole span should be annotated as one. For example, in "supervised and unsupervised methods", where we can also read "supervised methods and unsupervised methods", "supervised and unsupervised methods" is annotated as *one term*. Otherwise, split the string at the conjunction and annotate the conjuncts separately.

– Complex phrases containing *prepositions* can normally be split at those points where the prepositions are placed. Thus, for the text snippet ”automatic evaluation of machine translation and document summarization”, “automatic evaluation”, “machine translation”, and “document summarization” are annotated as separate terms.

  However, it is important to note that sometimes prepositions are a part of the term. Familiar examples are terms such as ”text to speech”, ”part of speech tagging”, ... Similarly, in relaxed paraphrases of normally denser terms (e.g., in “method for recognizing systematic patterns” instead of “pattern recognition method”), the whole sequence must be annotated as one term.

– In general, we aim at a *greedy annotation*. Some examples of what this means are given below:

  Example:

  * As explained earlier, given the string ”statistical machine translation method” which represents a single concept, the term is “statistical machine translation method”.
  * Given the string “supervised training data” in:

    > This paper presents a maximum entropy word alignment algorithm for Arabic-English based on supervised training data.

    annotate what is intended, (i.e., “supervised training data” as one term referring to ”manually annotated training data”). In other words, in this example *do not* annotate ”supervised training” and ”data” as two separate terms.
  * Given a nested term consisting of several concepts/terms (e.g., “maximum entropy word alignment algorithm”), annotate it as one term.
  * Given ad-hoc formations, (e.g., “suffix array-based data structure”), annotate them as a whole.
  * In the case of term paraphrases containing participles (e.g., “methods developed for spelling correction”), annotate them as a whole.

1.3 Decision Help

Table 2 provides decision help in difficult cases:

Category	Example	**Term?**
Domain high-level terms	algorithm, phrase, ...	YES
General scientific terms	experimental result, ...	NO
Contextual synonyms	e.g., "alignment" (when it is used instead of "word alignment"); "non-contiguous phrases" (to emphasise particular type of "phrases")	YES
paraphrases of more compact terms	e.g., "method for recognizing systematic patterns" instead of "pattern recognition method"	YES
Co-referencing items	e.g., "this analysis" or "this topic" vs. "semantic analysis" where "topic" and "analysis" refer to "semantic analysis" mentioned earlier in the text	NO

Table 2: Fine-grained semantic categories for annotation

1.4 Generic Nouns

1.4.1 Genrics that should be annotated

Here is an example of a generic that *should be annotated*:

> "The **data** from those recordings was used in a range of models ..."

In this example, "data" means "linguistic data" (semantic class: language resource) and thus can be viewed as a high-level domain term. Here, "data" should be annotated.

1.4.2 Generics that should not be annotated

Here are some examples of generic nouns that should *not* be annotated:

> "We describe an **implementation** of data-driven selection of emphatic facial displays for an embodied conversational agent in a dialogue system."

In this example, "implementation" cannot be mapped onto any of the classes described in Table 1 and, therefore, should not be annotated.

"This poster presents an **approach to spelling correction** . . . "
". . . results from **experiments with spelling correction** in Turkish."
". . . research into **methods for creating natural language text**."

In this example, the term candidate "approach to spelling correction" can be split by the preposition. "Approach" does not add specialised semantics to the term and should, therefore, not be annotated. The correct term is "spelling correction". The second snippet is similar: "experiments" is a general scientific term and does not add specialised semantics to the candidate. The correct term is "spelling correction". Example 3 is also similar to the first example in this quote: the correct term is "natural language text".

"After an overview of our **approach**, we present results . . . "

"Approach" is a general (scientific) term and should not be annotated.

2 Annotation Process

2.1 General Information

Annotators will receive annotation files in a simple XML format, as shown in Listing 1 in the appendix. Annotators are asked to *review* the automatic annotations in accordance with these guidelines. In particular, annotators are asked to:

- Add annotations within entity tags if they find a term has been missed the automatic annotator. The tag format is:

  `<entity>...</entity>`

- Remove incorrect entity tags.
- Correct the span of entity tags to fix the term length.

2.2 Files and File Exchange

Annotators will be able to pull annotation files from this GitHub repository: https://github.com/anetschka/Semeval-2018-Task-7-Entity-Annotation. For each batch of annotations, annotators will receive a folder that bears their first name. This folder will contain the files to be annotated.

Naming conventions:

- Files that need to be annotated will be named like this: (*number of abstracts*)_toAnnotate_(*annotator initials*).xml
- Annotators are asked to deliver their annotations in a new file named like this: (*number of abstracts*)_annotated_(*annotator initials*).xml

Annotators are asked to save the annotation files in the same folder as the raw files and then push the data back to the same GitHub repository.

References

Schumann, A.-K. and QasemiZadeh, B. (2015). The ACL RD-TEC Annotation Guideline. Published online: pars.ie/publications/papers/pre-prints/acl-rd-tec-guidelines-ver2.pdf. 1

Listing 1: Example of an automatically pre-annotated abstract.

```
<paper id="P02-1059"><title>
Supervised Ranking In Open-<entity>Domain</entity> <entity>Text
    ↪ Summarization</entity></title><abstract>
The <entity>paper</entity> proposes and empirically motivates an <entity>
    ↪ integration</entity> of supervised <entity>learning</entity> with
    ↪ unsupervised <entity>learning</entity> to <entity>deal</entity>
    ↪ with human <entity>biases</entity> in <entity>summarization</entity
    ↪ >. In particular, we explore the use of probabilistic <entity>
    ↪ decision tree</entity> within the <entity>clustering</entity> <
    ↪ entity>framework</entity> to account for the <entity>variation</
    ↪ entity> as well as <entity>regularity</entity> in human created <
    ↪ entity>summaries</entity>. The <entity>corpus</entity> of human
    ↪ created <entity>extracts</entity> is created from a <entity>
    ↪ newspaper</entity> <entity>corpus</entity> and used as a <entity>
    ↪ test set</entity>. We build probabilistic <entity>decision trees</
    ↪ entity> of different flavors and integrate each of them with the <
    ↪ entity>clustering</entity> <entity>framework</entity>. <entity>
    ↪ Experiments</entity> with the <entity>corpus</entity> demonstrate
    ↪ that the <entity>mixture</entity> of the two <entity>paradigms</
    ↪ entity> generally gives a significant boost in <entity>performance<
    ↪ /entity> compared to <entity>cases</entity> where either ofthe two
    ↪ is considered alone.
</abstract></paper>
```