The ACL RD-TEC

A Reference Dataset for the Evaluation of Automatic Term Recognition and Classification in Computational Linguistics

Annotation Guidelines

Behrang QasemiZadeh
behrangatoffice@gmail.com

Please read the following document before performing the annotation task. The annotator is required to understand the meaning of term, technology term, and invalid term before commencing the annotation task. A definition of each item is presented here.

1 Basic Definitions

Term: A term is a single a single token, a single word or a phrase consisting of several words/tokens and it characterizes a concept or a meaning in a technical domain. The Oxford Dictionary defines term as:

> 'a word or phrase used to describe a thing or to express a concept, specially in a particular kind of language or branch of study.'

According to ISO 1087-1(2000), a term is:

> 'a verbal designation of a general concept in a specific subject field.'

Linguistically, terms are *lexical units* and carry a special *meaning* in particular *contexts*. In the domain of computational linguistics, the following are examples of terms:

- ✓ lexicon
- ✓ dictionary
- ✓ corpus
- ✓ grammar formalism
- ✓ language resource
- ✓ natural language
- ✓ natural language processing
- ✓ machine translation
- ✓ statistical machine translation
- ✓ speech corpora

Technology Term: Among the terms, some of them refer to a technological concept. The Oxford Dictionary defines *technology* as:

1. The application of scientific knowledge for practical purposes, especially in industry: e.g. advances in computer technology;

2. Machinery and devices developed from scientific knowledge.

The Merriam-Webster Dictionary defines *technology* as:

1. A capability given by the practical application of knowledge, e.g. a car's fuel-saving technology;
2. A manner of accomplishing a task especially using technical processes, methods, or knowledge, e.g. new technologies for information storage;
3. Machinery and devices developed from scientific knowledge.

Last but not least, the Cambridge Dictionary defines *technology* as:

'the practical, especially industrial, use of scientific discoveries.'

Put simply, we categorize a term as a technology term if it refers to a concept that indicates a method or a process for accomplishing a task in order to fulfil a practical purpose. With these given definitions, among the list of terms that are itemized above, the following are assumed to be technology terms:

✓ natural language processing
✓ machine translation
✓ statistical machine translation.

As exemplified above, in computational linguistics, terms that indicate algorithms, methods, systems, practical approaches, frameworks, techniques, etc. form the category of technology terms.

Invalid Term: Invalid terms are those lexical units (words or phrases) that do not specify a key concept in the domain. These lexical units are generic words and phrases in the language. Any prepositional phrases, incomplete lexical units that signify terms, etc. are also considered as invalid terms. For example, in our task, the following are most probably not terms:

✗ engineering
✗ that the
✗ information access
✗ language dialogue
✗ for a statistical machine translation

See also the tips given at the end of document.

2 The Task

Given the definitions in Section 1 and using your knowledge and expertise, annotate the given set of candidate terms: mark them either as a term, a technology term, or an invalid term. Imagine a mind-map[1] of the topics in the computational linguistics. Would you like to see a given candidate term in this map (for instance, as visualized in Figure 1)? If the answer is yes, then this candidate term is probably marked as either a valid term or a technology term. Similarly, if you want to build a thesaurus/ontology of concepts in computational linguistics, will you incorporate a given candidate term in this thesaurus/ontology?

[1] http://en.wikipedia.org/wiki/Mind_map

In the provided tab-separated file, the first column shows the id of candidate terms, the second column represent terms' strings, and the third column is designated for the annotation mark. The annotator marks the third column with **1** for terms, **2** for technology terms and **0** for invalid terms. Also note :

- In order to decide whether a given candidate term is valid or invalid or a technology term, please refer to the ACL ARC corpus. The given text must confirm your understanding.
- Please note that technology terms are a subset of valid terms. This means that if you annotate a candidate term as a technology term, you automatically also annotate it as a valid term.
- If you annotate a candidate term (lexical unit) as invalid, this means that the given candidate term has never signified a key concept in the corpus. In contrast, if you annotate a candidate term as valid or a technology term, then it does not guarantee that all the occurrences of the term in the corpus are valid or a technology term.

Annotation Mark	Term Class
0	invalid term
1	domain term
2	technology term

Table 1: Annotation markers

Here are a few additional tips that can help you:[2]

- If a given candidate term is misspelt (e.g. miss-spelled!), as long as it is understandable, mark it similar to other terms; otherwise, it is marked as an invalid term. For instance, *word sense disamnbiguation* is not spelt correctly, but it may still be identifiable as the valid technology term *word sense disambiguation*.
- Ignore morphological/term variations. For example, in computational linguistics both *word sense disambiguator* and *word sense disambiguation* are valid technology terms. However, while the term *part-of-speech tagger* is a technology term, the term *part-of-speech tag* is just a term.
- If an abbreviation is so common that it is meaningful out of the context, e.g. *tf-idf*, then annotate it using the guideline given above. Otherwise, annotate it as an invalid term.
- We strongly recommend use of the concordance view of candidate terms using the Sketch Engine Corpus Query System available at https://the.sketchengine.co.uk/bonito/run.cgi/first_form?corpname=preloaded/aclarc_1. Alternatively, we can provide you with a desktop application. Also, you are allowed to use a web search for the given candidate term in, for example, Google Scholar and use the returned list of results in order to make your final decision.
- We specifically ask the annotator to bear in mind that although a term may be a valid technical term in a domain of expertise, it may not be a valid technology term. For example, in the domain of natural language processing, "language resource" and "WordNet" are valid technical terms. However, they are not technology terms. These terms do not refer to a process or method that can be used to address a

[2]Please feel free to discuss or challenge any of the given suggestions.

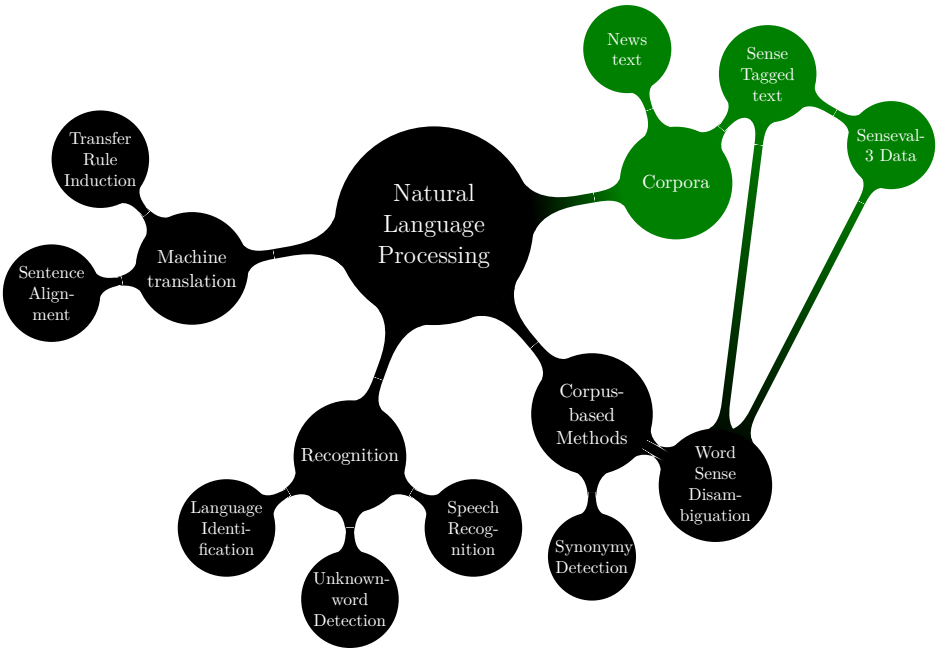

Figure 1: Example of a mind-map: black nodes represent technologies, while green nodes show other terms.

problem. Those terms that signify linguistic entities are perhaps more obvious examples; for instance, clitic, suffix, prefix, part-of-speech and syntax are valid terms in the domain, however, they are not technology terms. As a rule of thumb, read the term followed by the words technology, method, or algorithm, etc. and see if that term makes sense to you.

Thanks for your contribution. Your feedback can enhance this work. Please do not hesitate to contact me with any questions or suggestions. Results from a number of experiments and methods for preparing the raw data can be found in [1, 2, 3].

References

[1] Behrang Q. Zadeh and Siegfried Handschuh. Evaluation of technology term recognition with random indexing. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, May 2014. European Language Resources Association. ACL Anthology Identifier: L14-1703.

[2] Behrang Q. Zadeh and Siegfried Handschuh. The ACL RD-TEC: A dataset for benchmarking terminology extraction and classification in computational linguistics. In Patrick Drouin, Natalia Grabar, Thierry Hamon, and Kyo Kageura, editors, *COLING 2014: Computerm 2014: 4th International Workshop on Computational Terminology: Proceedings of the Workshop*, pages 52–63, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[3] Behrang QasemiZadeh and Siegfried Handschuh. Investigating context parameters in technology term recognition. In Adam Meyers, Yifan He, and Ralph Grishman, editors, *Proceedings of the COLING Workshop on Synchronic and Diachronic Approaches to Analyzing Technical Language (SADAATL 2014)*, pages 1–10. Association for Computational Linguistics and Dublin City University, 2014.