SemEval-3 Task Proposal

Semantic Compositionality in Context

Organizers

Chris Biemann (UKP Lab, Technische Universität Darmstadt)
Eugenie Giesbrecht (Forschungszentrum Informatik an der Universität Karlsruhe)
Torsten Zesch (UKP Lab, Technische Universität Darmstadt)

Task Description

An interesting sub-problem of semantic compositionality is to decide whether a phrase is used in its literal or figurative meaning in a given context. For example “big picture” might be used literally as in

> “Click here for a *bigger picture*”

or figuratively as in

> “To solve this problem, you have to look at the *bigger picture*.”

Another example is “old school” which can also be used literally or figuratively.

> “He will go down in history as one of the *old school*, a true gentlemen.”
> “During the 1970’s the hall of the *old school* was converted into the library.”

Being able to detect whether a phrase is used literally or figuratively is important for a wide range of tasks in Natural Language Processing, e.g. document retrieval, clustering and classification, question answering, query expansion, word similarity, synonym extraction, relation extraction, or textual advertisement matching in search engines.

Participants of this task will be provided with a list of target phrases together with real usage examples sampled from the large-scale freely available WaCky (Baroni et al., 2009) corpora.[1] For each usage example, the task is to make a binary decision whether the target phrase is used literally or figuratively in this context.

Expected Scientific Benefit

The range of applicable methods for this task is deliberately **not** limited to a certain branch of methods (e.g. distributional models of semantic compositionality), as it can be tackled from different directions, and we expect a great deal of the expected scientific benefit to lie in the comparison of very different approaches to solve this challenging task, as well as how different approaches can be combined.

Participants might make use of pre-fabricated lists of phrases annotated with their probability of being used figuratively from publicly available sources. They might use selectional preferences or deep semantic parsing for deciding whether a phrase might be used figuratively in most cases (e.g. “kick the bucket”). Assessing how well the phrase suits its context might be tackled using measures of semantic relatedness as well as distributional models learned from the underlying corpus. The task may also be of interest to the related research fields of metaphor detection and idiom identification.

Targeted Languages

The task will contain subtasks for the following languages:

- English
- German

[1] http://wacky.sslmit.unibo.it/

Relation to Outline Proposal 12

- The task as described in the original proposal has been run by some of us as a shared task for DiSCo-2011: Workshop on Distributional Semantics and Compositionality, in conjunction with ACL 2011. We decided to elaborate on that and focus on a particular phenomenon we found when preparing the DiSCo task.
- The task is not longer focused on distributional approaches alone. We expect to attract a wider range of participants and also to gain a higher scientific impact from the comparison of different approaches.
- Instead of focusing on the compositionality of a phrase per se, we put that question into context by providing real usage examples. For example, a phrase that is used in a non-compositional sense in almost all cases might still be used compositional in a certain context. We are interested in assessing to what extent the current state-of-the-art is able to tackle this challenging problem.
- Due to a change in the group of organizers, we cannot offer the Italian subtask any more.

Data

For each language, the organizers will extract about 100 target phrases and 1000 contexts from the WaCky corpus. Each context will be annotated using the binary classification scheme (literally vs. figuratively) by native speakers of the respective language. The annotators will be recruited through web-based services such as Amazon Mechanical Turk or similar tools. The organizers have previously carried out similar annotation work before in context of the DiSCo 2011[2] shared task and other data acquisition projects.

The provided context items will be already lemmatized and POS tagged, as provided by the WaCky initiative. Participants, whose approaches make use of either POS tags or lemmas, are strongly encouraged to use this shared preprocessing to ensure the highest possible comparability of results. Additionally, it may considerably reduce the work-load on the participants' side. To further lower the boundaries of entering the task, we will provide UIMA-based components for working with the data.

The data will be split into training, validation and test sets. We define two different subsets that will be scored separately and in combination:
1. One subset of target phrases will be accompanied by a large number of contexts, which allows learning single classifiers for each phrase. The validation/test sets will only contain phrases that occur in the training set. This is comparable to the lexical sample task, which encourages the participation of supervised systems.
2. The other subset will contain target phrases together with a smaller number of usage contexts, probably favoring unsupervised approaches. The validation/test sets will contain new target phrases. This is comparable to the all-word task. Here, systems must grasp a notion of literacy vs. idiomatic use in general, without training classifiers for each phrase.

The training and validation portions will be made available to the participants, together with a scoring infrastructure. For the challenge, participants submit their system's output on the test sets to the task organizers, who score the systems and provide the official scores.

Corpus sharing

Since our task is focusing on a particular phenomenon – phrases that can be understood literally or idiomatic – we do not think that we can share resources with any of the other tasks beyond using the same (large) background corpus. We will gladly use another corpus besides WaCky, as long as its size is comparable and it comes preprocessed with POS and lemmatization.

Evaluation Criteria

System performance will be measures in terms of precision/recall/F1 of the classifier. Systems are encouraged to participate in both subtasks, but submissions to single subtasks are accepted. For systems tackling both subtasks or both languages combined scores reflecting the overall performance will be computed.

[2] http://disco2011.fzi.de/

Besides the overall performance score on the full test set, we will provide sub results for phrases already seen in the training data as well as for the new target phrases.

Availability of Resources

The large-scale WaCky (Baroni et al., 2009) corpora are freely available. As only very small annotated samples from freely available web documents are used, and the original source is provided, they can be distributed without violating copyrights. The size of the underlying corpora allows for reliable distributional models to be trained.

Resources required to prepare the task

For preparaton of the task, we need large corpus resources and a list of candidate phrases. For corpora, we will use the freely available WaCky corpora. Lists of candidate phrases can be found in Wiktionary or other dictionary resources. Thus, we have everything at hand and could start preparing the data immediately.

References

M. Baroni, S. Bernardini, A. Ferraresi and E. Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation 43 (3): 209-226.