SemEval-2010 Task 9
Noun Compound Interpretation Using Paraphrasing Verbs

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz and Tony Veale

1. Task Description

- This task requires systems to estimate the goodness of verbal and verb-preposition paraphrases of English compound nouns.
- For each compound in the dataset, systems are provided with a set of possible interpretations and should rate the goodness of each interpretation in accordance with the ratings of human subjects.
- This is new task for SemEval; we build on previous work by Nakov (2008) and Butnariu and Veale (2008).

2. Semantics of Noun Compounds

- Noun compounds can (and do) express a great variety of semantic relations between their constituents. How best to model this variety is an open question.
- Inventory assumption: compound meaning can be captured by a small set of relational categories, e.g., Levi (1978):

  chocolate bar → HAVE
  fruit tree → HAVE?/MAKE?
  sleeping pill → FOR
  headache pill → FOR

- Problem: Categories can conflate heterogeneous meanings, and meanings can be ambiguous as to the correct category.
- Alternative model: capture compound meaning through paraphrases. Lauer (1995) uses prepositional paraphrases, but these are too restrictive.
- Following Nakov (2008) we use paraphrases of the form *N that Verb N* or *N that Verb Preposition N*.
- Instead of a single paraphrase per compound we assume a distribution over likely and unlikely paraphrases. For example:

  fruit tree → *bear*(20); *produce*(16); *grow*(15); *have*(6); *give*(4); *provide*(3); *develop*(2); *supply*(2); *make*(2); *hold*(1); *contain*(1); *bare*(1); *be laden with*(1); *be grown for*(1); *be filled with*(1); *be made from*(1); *bloom*(1)...

3. Data Collection

- Compound paraphrases will be collected from human subjects using the Amazon Mechanical Turk service (www.mturk.com).
- Standard of MTurk annotators is high (Snow et al., 2008). Large quantities of annotations can be collected quickly and at low cost compared to traditional methods.
- Methodology adapted from Nakov (2008):

  (1) All subjects must pass a simple preliminary test to check their language competence and general answer quality.

  (2) Each MTurk Human Intelligence Task (HIT) involves giving three or more paraphrases for each of five compounds.

  (3) Each compound is paraphrased by multiple subjects; the responses are collated to give a distribution over paraphrases.

- "Large number" assumption: the most frequently given paraphrases correspond to probable interpretations, while unpopular paraphrases are unlikely interpretations or annotation noise.

4. Dataset and Evaluation

- Training/development dataset consisting of paraphrases for 250 compounds previously compiled by Nakov (2008).
- New test dataset of 300 compounds each paraphrased by ~100 MTurk users.
- Official evaluation measure is the average cosine similarity between the system scores for the interpretations of a compound and the frequency distribution provided by the annotators.
- We will also report other measures (e.g., Spearman correlation) and a qualitative analysis.

5. More Information

- Task web page: http://groups.google.com/group/semeval-2010-noun-compound-interpretation-using-verbs
- Contact: Preslav Nakov (nakov@comp.nus.edu.sg)