CoNLL Shared Task 2010 Proposal: Reference Resolution to Wikipedia

September 21, 2009

1 Organizing Committee

1. Lev Ratinov (University of Illinois at Urbana-Champaign), contact person.
2. Doug Downey (Northwestern University)
3. Dan Roth (University of Illinois at Urbana-Champaign)

2 Task Description

A fundamental problem in natural language processing involves mapping textual strings to their referant entries in an external knolwedge base. Wikipedia has rapidly become the canonical knowledge base for this task, due to its unprecedented breadth. We refer to the task of unambiguously identifying Wikipedia concepts mentioned in a given text as *Reference Resolution to Wikipedia* (RRW). RRW has recently been shown to be valuable across a wide spectrum of NLP tasks including text classification [6], measuring semantic similarity between texts [7], information retrieval [3, 1, 2], word sense disambiguation [10, 4], cross-document coreference resolution [5, 9], and other tasks [11, 8].

Excitement about RRW has resulted in a variety of distinct approaches, developed by numerous research groups. Due to differences in the specific task formulation and annotated corpora employed, the approaches are difficult to compare. This proposal aims to address these challenges by providing a unified evaluation suite for RRW as the CoNLL 2010 Shared Task. Below, we begin by describing previous variants of the RRW task, and then detail our proposed task and data collection methodology.

2.1 Previous Formulations and Data Sets

The RRW task has been formulated in two distinct ways in recent research. In the first task formulation, a system is given an input text and must output relevant Wikipedia concepts that are explicitly or implicitly referred to in the input text, without necessarily binding the concepts to specific mentions (or surface forms) in the text [6, 7, 3, 5]. In the second task formulation, the system must link specific text mentions to their Wikipedia referants [1, 2, 10, 11, 8]. For example, given the input text "*I went to Broadway to watch Chicago on Saturday, 7 June 1975.*", candidate output for the first formulation might include:
(1) Musical_theatre
(2)Broadway_theatre
(3) Bob_Fosse
(4)List_of_the_100_Longest-Running_Broadway_shows.
For the second formulation, the output might be: "*I went to <Broadway url=http://en.wikipedia.org/wiki/Broadway_theatre> to watch <Chicago url=http://en.wikipedia.org/wiki/Chicago_(musical)> on Saturday, 7 June 1975.*".

The latter task is referred to as *Wikification* in the literature [10], and has the advantage of being more straightforward to define: the entities *explicitly* referenced in a text are typically a more easily-identified subset of all those concepts "relevant" to a text. For this reason, our proposed task will focus on identifying concepts explicitly mentioned in an input text.

However, even within the Wikification formulation, previous work varies along multiple dimensions. Previous efforts differ in the set of references they attempt to identify (e.g. some identify only named entities, others attempt to find *all* phrases with Wikipedia referants, and still others attempt to mimick the non-redundant link structure of Wikipedia). Further, evaluation corpora varies from high-quality newswire text, perhaps assuming gold-standard named entities have been identified, to general Web text. Because the difficulty of the disambiguation task changes markedly depending on these design decisions, precision and recall on RRW vary widely and comparisons between different approaches are rarely meaningful.

The variability described above comes from an aim to use RRW to address different applications (e.g. allowing faceted search for named entities, or marking-up Wikipedia with new, helpful hyperlinks). The goal of our shared task is not to address a specific application, but instead to focus on the common RRW sub-problem

shared about these applications. Thus, we do not consider *locating* the strings to map to Wikipedia as part of our RRW task, as previous Wikification efforts often do. Instead, our proposed data set provides these strings as input data for two important reference sets discussed above (entities and non-entities).

2.2 Formal Proposed Task

We formulate the RRW shared as follows. The system is given a input text, and an expression within the text to be mapped to to Wikipedia. The system is to either output the corresponding Wikipedia concept, or to output that no such concept exists.

Sample problem/answer pairs are[1]:
Q: "[Florida] authorities have requested federal assistance in locating a man sought for questioning in the deaths of his wife and five children, as he is believed to be in Haiti, officials said Monday."
A: http://en.wikipedia.org/wiki/Florida

Q: "[Florida authorities] have requested federal assistance in locating a man sought for questioning in the deaths of his wife and five children, as he is believed to be in Haiti, officials said Monday."
A: http://en.wikipedia.org/wiki/Police

Q:"Florida authorities have requested [federal] assistance in locating a man sought for questioning in the deaths of his wife and five children, as he is believed to be in Haiti, officials said Monday."
A: http://en.wikipedia.org/wiki/Federal_Bureau_of_Investigation

Q:"Florida authorities have requested federal [assistance] in locating a man sought for questioning in the deaths of his wife and five children, as he is believed to be in Haiti, officials said Monday."
A: null

Our textual corpus will consist of 3818 articles taken from NYT newswire data, along with blog data acquired independently from RSS feeds. Both datasets will be freely available. We will employ two different approaches for finding candidate expressions. In approach A, we will identify all substrings in the document which refer to some Wikipedia concept (similar to [8]), and choose among these at random. In approach B, we will have human annotators read the document and choose the five textual strings they would most like to investigate to better understand the document. For both approaches, human annotations of the ground truth mapping to Wikipedia will be obtained using Amazon's Mechanical Turk (our previous experience has shown this approach to be effective for RRW).

The eight different combinations of data sets (newswire text and blog text), approaches A and B, and expression types to be linked (named entities and non-entities) will serve as eight separate evaluation tracks. We believe these tracks exercise the key dimensions in corpora and reference set selection required for varying applications, while providing a unified and standardized evaluation process. Systems will be evaluated in terms of precision and recall on each track, independently. We will provide the necessary evaluation software, a baseline system, and a preprocessed version of Wikipedia, along with an interface similar to the one described in http://wikipedia-miner.sourceforge.net.

References

[1] R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL'06*, 2006.

[2] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *EMNLP-CoNLL'07*, 2007.

[3] O. Egozi, E. Gabrilovich, and S. Markovitch. Concept-based feature generation and selection for information retrieval. In *AAAI'08*, 2008.

[4] A. Fader, S. Soderland, and O. Etzioni. Scaling Wikipedia-based Named Entity Disambiguation to Arbitrary Web Text. In *WikiAI (IJCAI workshop)*, 2009.

[5] T. Finin, Z. Syed, J. Mayfield, P. McNamee, and C. Piatko. Using wikitology for cross-document entity coreference resolution. In *AAAI Spring Symposium on Learning by Reading and Learning to Read*, 2009.

[6] E. Gabrilovich and S. Markovitch. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with. In *AAAI'06*, 2006.

[7] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI'07*, 2007.

[8] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of wikipedia entities in web text. In *KDD'09*, 2009.

[9] J. Mayfield and et al. Cross-Document Coreference Resolution: A Key Technology for Learning by Reading. In *Proceedings of the AAAI 2009 Spring Symposium on Learning by Reading and Learning to Read*. AAAI Press, March 2009.

[10] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM'07*, 2007.

[11] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM'08*, 2008.

[1]The text is taken from http://www.cnn.com/2009/CRIME/09/21/florida.family.dead/index.html