**Entity Linking**

**Principles of our Entity Linking system:**
- Simple system
- Context independent dictionary that maps entity mentions to candidate Wikipedia articles
- Context dependent heuristic disambiguation using NER and coreference
- NIL as matches of entities in Wikipedia outside the KB

**Previous work:**
- For TAC-KBP 2009 we developed two different disambiguation components
  - A supervised disambiguation system, based on Word Sense Disambiguation techniques.
  - A knowledge-based system, using overlap with test of KB and Wikipedia articles.

**Dictionary Construction**

This is the centerpiece of our system, and includes all potential string to entity pairs.
- include all KB entities
- augment all articles in Wikipedia dump
- build equivalence classes using redirects and canonicalization
- choose representatives article, KB entity if possible

**String extraction:**

String were collected as follows, producing three dictionaries:
- titles of articles, either literally or after deleting (...) 
- titles of disambiguation pages (EXACT)
- performing lower case normalization (LNRM)

**String to entity mapping:**

String to entity maps are weighted using occurrences of the string as anchor text of the a link to the entity in:
- w: Wikipedia text
- W: The Web

In addition we also explored Google rank for string in en.wikipedia.org (GOOG).

**Heuristic Disambiguation**

Simple heuristic based approach that attempts to disambiguate an entity mention by identifying other possible mentions of the same entity in the text. The intuition is as follows:
- Issue mentions are more ambiguous than others
- Examples: synonyms such as “ABC.”
- Other mentions of the same entity in the document are less ambiguous
- Example: “American Broadcasting Company.”
- To keep things simple, we assume that the longest entity mention is least ambiguous

**Disambiguation Rules:**

Find the set of possible mentions for a given entity mention
- Run NER on document text to find all entity mentions
- For all occurrences of the target entity string, find the longest NER-chunk it was part of.
- Use coref to find all entity names that are coreferent with the target entity
- Find all matches of Wikipedia titles that target mentions can refer to in document
- Identify Wikipedia title based on longest matching entity string

**Simple System:**

The mention alone, without the surrounding context.
- The last occurrence of the mention in the text, with a span of 25 tokens to the right of the right of the mention.
- The set of missing of all matching mentions, and their 25 token spans, in the text.
- A window of 1,000 tokens around the last mention in the text.
- Preliminary results on a development set showed that using the concatenation of all occurrences of the mention and their 25 word contexts performed best, and that this was used for test set queries.

**Future Work:**

We plan to combine the heuristic disambiguation system with the robust system we developed for TAC-KBP 2009.
- Heuristic setting — vote using inverse of rank of GOOG, the cascaded dictionary,
- Optimized set — a linear combination of positive weights applied to the scores from many components.