Pump Priming: Semantic Information Extraction using Integrated Probabilistic Techniques

Wray Buntine[1], Helsinki Institute for Information Technology
Cyril Goutte and Eric Gaussier, XEROX RC Europe
Marko Grobelnik and Dunja Mladenic, J. Stefan Institute

Institut "Jožef Stefan"

[1]Now at NICTA.

Goals

- *Semantic information extraction* is an important task.
- We *claim* it requires
  - integrated methods with a probabilistic representation,
  - and use of large scale semantic/linguistic resources.
- The project will use the public domain Wikipedia as a linguistic and (rough) semantic resource.
- We will explore semantic integration across the language processing chain.

Wikipedia

- Getting to over two million articles, increasing rapidly, each on a GNU Free Documentation License.
- Early input included imports from the 1911 Encyclopedia Britannica and government fact books. Very popular with fans of TV series. movies, comics and such.
- Founder, Jimmy Wales dabbled in pornography prior to his entry into the "free knowledge" business, and now branching out into search.
- Items authored freely, so corporations, government agencies and political parties actively "scrub" their content. Claims of bias and inaccuracies are predictably leveled.
- Text is good quality, large variety of content and structural styles, and riddled somewhat erratically with tags (as links) and topics (as categories).

Information Extraction and Natural Language Processing

- Many large scale users of IE and NLP software believe that well-developed hand-coded systems are superior to statistical NLP systems *due to the ability to be robust outside the domain of development*.
- On custom domains where appropriately marked up linguistic resources exist, statistical NLP systems can be better.
- Getting adequate quantities of marked up linguistic resources therefore is a problem.
- Alternatively, one can resort to:
  - Alternative linguistic resources.
  - Use of semi-supervised or active learning approaches.
  - Robustification, for instance in choice and development of features.

Proposing Links

Problem Definition: given a candidate term in a Wikipedia page, propose the most appropriate semantic link (*i.e.*, a URL to another Wikipedia page) that explains the term).

> *The page ACE_inhibitor has incoming anchor text including linguistic variants of "ACE inhibitor". "angiotensin converting enzyme inhibitor" and "inhibitor". Outgoing anchor text includes "volume" to stroke_volume and "fatigue" to Fatigue_(physical).*

Issues:

- The term can have exact or approximate matches with existing anchor text.
- A term can have several link candidates for such a match (polysemy).

Proposing Links, cont.

- An automated system developed at Xerox in 2006.
- Context of each term (5 words on each side) is used to predict the DMOZ category for the context.
- Each document is assigned a DMOZ category using a probabilistic generative model.
- Thus candidate documents matching the term (in their anchor text) can be chosen based on category match.
- Method shows promise.

Other Useful Tools

- Some support tools developed at HIIT in 2006 to make Wikipedia content more useful and more accurately marked-up.
- Distinction between whether pages are for proper names, terms, or are ambiguous entries for *disambiguation*.
- Typing (*e.g.*, a name as a *person*, *company*, *place*) done patchily in Wikipedia so tools to recognise whether pages are for people, companies, and time events where developed.
- Developing cross-lingual resources from the page links.
- A search engine in Lucene.

These last two tasks moved into the SMART project.

Named Entity Resources

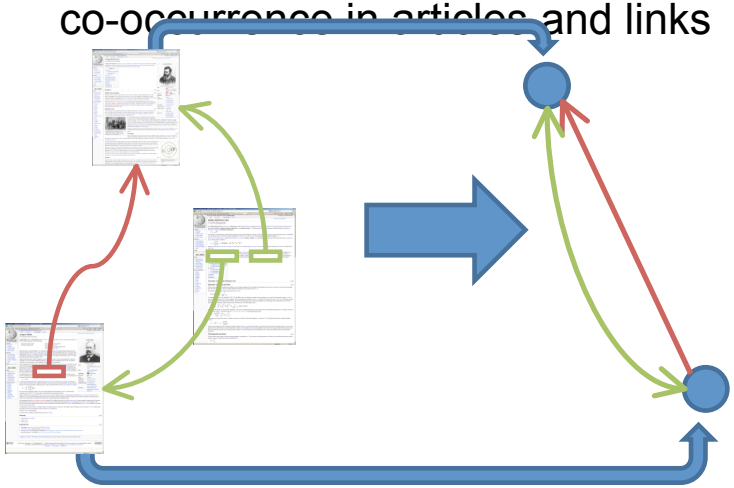
- Wikipedia pages by themselves do not provide good tagged metadata.
  - Only the first occurrence of the name in the text is usually tagged.
  - Often a reference is used, rather than a name, *e.g.*, the anchor text "the President" is tagged with a URL to *William_Jefferson_Clinton*.
  - Only some pages correspond to unambiguous named entities.
- Thus we need modified learning algorithms:
  - to deal with partially tagged named entity data (*e.g.* semi-supervised conditional random field),
  - and to address the "reference" versus "name variant" distinction problem.

Extraction of People and Places

- Heuristics approaches
  - Using lists
    - For people: “List of People”, “List of Slovenian computer scientists”, …
    - For places: “List of Countries”, …
    - For organizations: “List of Universities”, …
  - Using Infoboxes
    - Templates for providing standard information to the article
    - Types of infoboxes were manually mapped to instance type
  - Using article structure:
    - Most articles about people have born/died dates at the beginning
    - Many places have coordinates at the top of the page
- Learning approach
  - **Bootstrap out of articles extracted using heuristics**
  - Used linear SVM as classifier
  - Bag-of-words vectors generated from articles enriched with extra features
  - Categories:
    - Person, Location, Organization, Other

Historical Social Network and Events

- People, locations and organizations were connected based on co-occurrence in articles and links



- Output is a big social-network placed in time (born- died dates) and space (locations from people's articles)

Extraction of events:

- Only for instances of people
- Subset of sentences extracted from articles.
- Example for **Jozef Stefan:**

1845	They recommended that he continue his schooling, so in 1845 he went to Klagenfurt gymnasium.
1848	He experienced the revolutionary year of 1848, as a thirteen-year-old boy, which inspired him to be sympathetic toward Slovene literary production.
1853	He left for Vienna in 1853 to study mathematics and physics.
1857	Stefan then graduated in mathematics and physics at the University of Vienna in 1857.
1866	He taught physics at the University of Vienna, was Director of the Physical Institute from 1866, Vice-President of the Vienna Academy of Sciences and member of several scientific institutions in Europe.
1884	In 1884 the law derived theoretically in the framework of thermodynamics by his student Ludwig Boltzmann and hence known as the Stefan-Boltzmann law.

Topic Models and Compounds

- A number of techniques seek to integrate a compound detection capability with a topic modelling capability along the lines of LDA/DCA.
- Some models replace the word probabilities inside the model with n-gram based word probabilities. So, "Its a German Shepherd" modelled as
$$p(Its|t_1)p(a|t_2)p(German|t_3)p(Shepherd|t_4, German).$$
  - *While a good model for perplexity scores, and simple conceptually, compounds are not directly modelled.*
- An alternative approach would group select word sequences (*i.e.* into compounds). So
$$p(Its|t_1)p(a|t_2)p(GermanShepherd|t_3).$$
  - *This requires compound creation and deletion operators and flexible topic-word model priors. More difficult!*

Conclusion

- The Historical Social Network is a great candidate for additional study.
- Wikipedia has become a significant resource for named entity recognition, but more work is needed.
- Tools to support tagging and semantic cleaning of the Wikipedia are emerging, and prove important.
- Alternative topic models are in development.
- A number of other efforts where pursued that didn't produce successful results.
- Spinoff of results into SMART project (an FP6 STREP with IST).