Scalable Semantic Annotation of Text Using Lexical and Web Resources

Elias Zavitsanos
Web: http://users.iit.demokritos.gr/~izavits/
e-mail: izavits@iit.demokritos.gr

National Centre for Scientific Research "Demokritos"
Institute of Informatics and Telecommunications

George Tsatsaronis
Web: http://www.idi.ntnu.no/~gbt/
e-mail: gbt@idi.ntnu.no

NTNU
Norwegian University of Science and Technology

Iraklis Varlamis
Web: http://www.dit.hua.gr/~varlamis/
e-mail: varlamis@hua.gr

HAROKOPION UNIVERSITY

Georgios Paliouras
Web: http://users.iit.demokritos.gr/~paliourg
e-mail: paliourg@iit.demokritos.gr

National Centre for Scientific Research "Demokritos"
Institute of Informatics and Telecommunications

Presentation Layout

- Introduction and Motivation
- Contributions
- Related Work: Automated Annotation of Text Documents
- Method of KDTA
- Experimental Evaluation
- Conclusions and Future Directions

Introduction and Motivation

- (Semi-)Automated annotation of text enables:
  - Transfer of useful information from text documents to the ontology
  - Key step to proceed with (semi-)automated knowledge extraction
  - Fundamental technology for intelligent Web applications
- Ontologies pertain to specific domains
  - Most approaches require training data to annotate new documents
  - Supervision from domain experts
  - Requires much time and effort
- Need: Methods that may annotate plain text given a domain ontology, without training
  - Use of lexical and/or Web resources
  - Advanced NLP techniques (WSD)
  - Use of measures that capture similarity between text segments and terms (i.e., measures of semantic relatedness and/or similarity)
  - All these should allow fast execution for on-line annotation

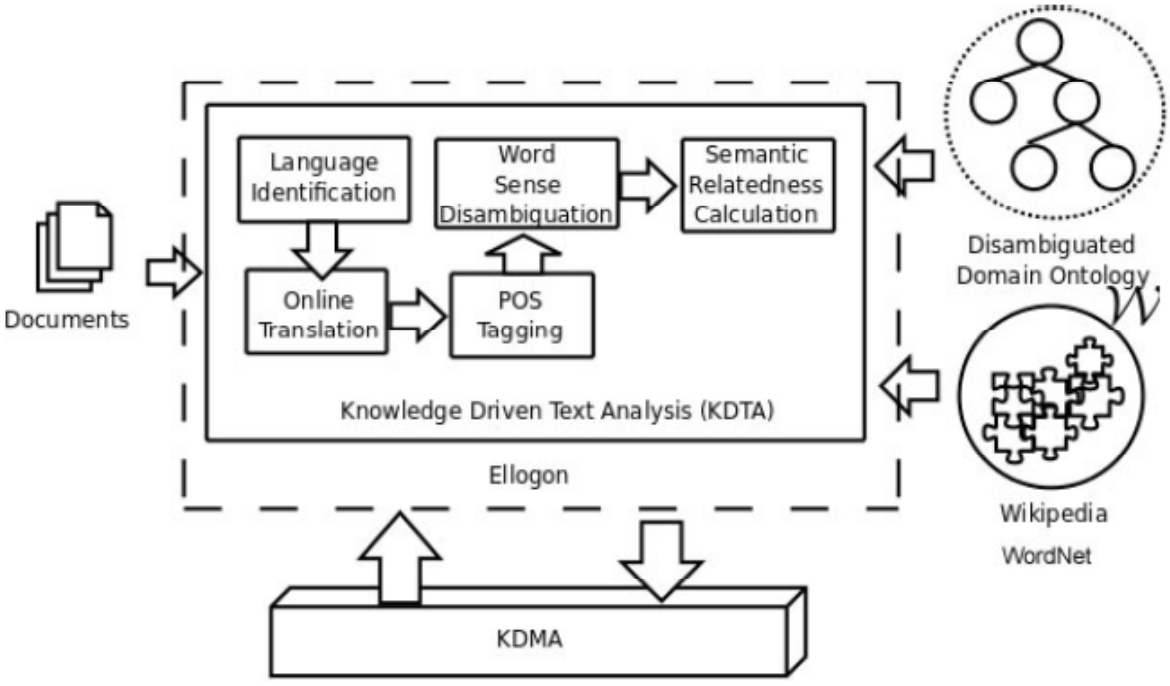
Contributions

- A novel method for semantic annotation of plain text with ontology concepts
  - Totally unsupervised
  - Using Lexical and Web Resources
    - WordNet
    - Wikipedia
- Thorough experimental evaluation
  - Environmental and Medical domains
  - Study the effect of several method components (i.e., WSD)
- Testing of the method for on-line document annotation, under the framework of the CASAM project (Knowledge-driven Text Analysis module / KDTA)

Related Work

- *PANKOW* and *C-PANKOW* tools [Cimiano et al., 2005]
  - Very slow, bounded by the use of Google API
  - Named-entity annotation
  - Restricted to certain types of named enities
  - High precision, but very limited recall
- [Ding and Embley, 2006]
  - Automated semantic annotation of Web pages
  - Domain expert needed to formalize the semantics of the domain
  - Data-extraction ontologies, used to avoid heuristics of IE techniques
- *Ontea* System [Laclavik et al., 2006]
  - Predefined domain specific regular expression patterns
  - Domain ontology needs to incorporate special ontology extension used by *Ontea*
- Other tools (*CREAM*, *Magpie*)
  - Provide useful visualization for manual annotation of documents

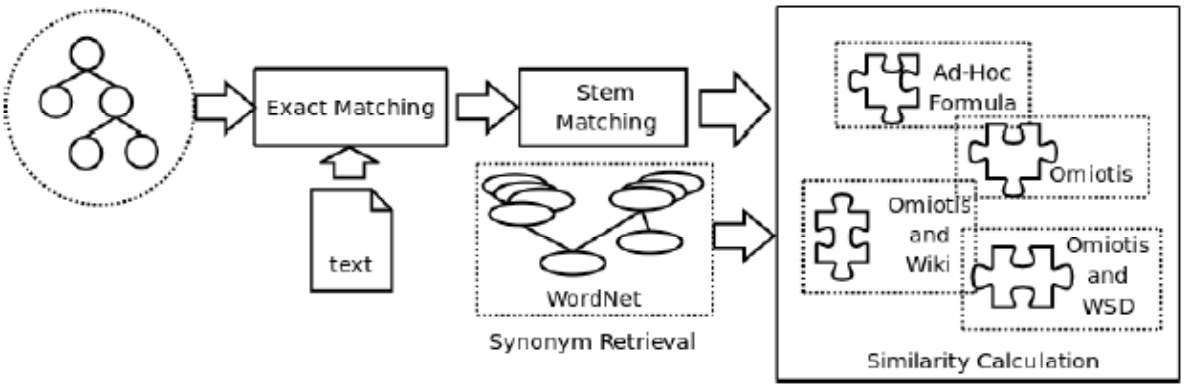
General Overview of the Method



Steps of the method

- Pre-processing
  - Language identification and online translation (English)
  - Part-of-speech tagging (provides useful information for the disambiguation process)
  - Named entity recognition
  - Focus on nouns or noun-phrases
  - Word Sense Disambiguation using a PageRank-based method (unsupervised)
- Annotation with Ontology Concepts
  - Exact matching
  - Stem Matching
  - Semantic relatedness/similarity matching

Overview of the Annotation Step



Measuring Semantic Relatedness/Similarity

- Baseline Method

$$SD = \frac{1}{NS * \frac{\log CS}{\log NS}}$$

- Annotation with the Omiotis measure of semantic relatedness

$$SR(s_i, s_j) = \max(SCM(s_i, s_j) \bullet SPE(s_i, s_j))$$

$$SR(t_i, t_j) = \max(SR(s_i, s_j))$$

- Annotation with Omiotis, supplemented by the Wikipedia-based measure of Milne and Witten
  - Use the Wikipedia-based measure when Omiotis cannot be computed

Embedding WSD

- During the computation of semantic relatedness, if the correct senses of the words are known:

$$SR(t_i, t_j) = SR(s_i, s_j) = \max(SCM(s_i, s_j) \bullet SPE(s_i, s_j))$$

- where: s_i, s_j are found by the used WSD algorithm to disambiguate the words t_i, and t_j respectively

Implementation and Complexity

- Omiotis has indexed all pair-wise synset-to-synset relatedness values
- Wikipedia is installed locally, and the measure of Milne and Witten is very fast
- Overall implementation through the Ellogon platform
  - Annotation measures are components and can be used/substituted very easily
- Flexible architecture, that allows Ellogon modules easy re-use and replacement

Evaluation in Environmental Domain

- 51 documents provided by the LUSA agency
- Corresponding ontology comprises 230 concepts
  - Environmental concepts
    - e.g., “Wind”, “Water”, “Solar Energy”
  - Technological concepts
    - e.g., “Media Equipment”, “Car”, “Building”
  - Entities
    - e.g., “Person”, “Profession Name”
- Manual annotations were provided, with ontology concepts, as a means of correct annotations
- Measure: macro-averaged Precision, Recall, and F-Measure of each annotation method, using the manual annotations as gold standard

Results in LUSA

	Baseline	Omiotis	Omiotis&WSD	Omiotis&Wiki
Macro Avg. Precision	0.73	0.51	0.54	0.51
Macro Avg. Recall	0.76	0.57	0.55	0.58
Macro Avg. Fmeasure	0.73	0.49	0.51	0.50

- The ad-hoc computation of relatedness seems to work better in this case
  - Most manual annotations occur from words which exist in the exact synsets' list of the initial term, as found in WN.
- WSD does not increase much the performance
  - Low ambiguity of words (nouns)
  - Use of unsupervised WSD method (est. 61-63% in Senseval)
- The supplementary use of the Wiki-based measure improves also little
  - Better normalization of Omiotis and Wiki values

Evaluation in the Molecular Biology Domain

- GENIA ontology comprising 49 concepts
- 2000 Medline abstracts, annotated with Genia concepts
- More difficult than LUSA
- Very specific medical terms
- WordNet and Wiki offer really small coverage in this domain

Results in the GENIA corpus

	Baseline	Omiotis	Omiotis&Wiki
Macro Avg. Precision	0.72	0.30	0.36
Macro Avg. Recall	0.08	0.26	0.27
Macro Avg. Fmeasure	0.15	0.28	0.31

- Semantic relatedness performs much better than ad-hoc method
  - Annotated terms are distant from the respective ontology terms, which baseline fails to capture
- Increase in recall by almost 20 p.p., and F-Measure twice as good
- Both WordNet and Wiki cannot cover such specific terms
  - UMLS might be more proper in this case

Conclusions

- A totally unsupervised method for free text annotation
  - F-Measure up to 73% in the LUSA case study
- The proposed architecture is flexible, components can be easily re-used and replaced
- Measures of semantic relatedness seem promising, especially to boost the recall of the overall performance
- Word sense disambiguation seems to improve very little, but this might not be the case with the other domains which carry higher ambiguity
- Combination of WordNet and Wiki certainly offers wider coverage

Future Directions

- Better combination of Omiotis and Wiki-based measure (normalization)
- Use of cut-off values for the relatedness measures
- Domain-biased WSD
- Embed more measures of semantic relatedness, and ensemble
- Relatedness as probability for the recommendations

Questions

Thank you very much for your attention!

Questions/Comments?