

Towards the Self-Annotating Web by Philipp Cimiano, et al.

Presented by Will Lee
wlee1@uiuc.edu

September 15, 2004

Motivation

- Semantic Web requires deeper understanding behind the words
- Meta-data hard to come by without a lot of manual labor
- Desirable to generate the meta-data (annotations) in an unsupervised way

Methods

1. Extract text from web page

Methods

1. Extract text from web page
2. Use part of speech (POS) tagging to find the proper nouns (instances)

Methods

1. Extract text from web page
2. Use part of speech (POS) tagging to find the proper nouns (instances)
3. Come up with a list of ontology concepts (concepts)

Methods

1. Extract text from web page
2. Use part of speech (POS) tagging to find the proper nouns (instances)
3. Come up with a list of ontology concepts (concepts)
4. Generate a set of hypothesis phrases using *instance* × *concept* × *pattern*. For example:

Methods

1. Extract text from web page
2. Use part of speech (POS) tagging to find the proper nouns (instances)
3. Come up with a list of ontology concepts (concepts)
4. Generate a set of hypothesis phrases using *instance* × *concept* × *pattern*. For example:
 - (a) Pattern (to be described next): e.g. “<instance> is a <concept>”

Methods

1. Extract text from web page
2. Use part of speech (POS) tagging to find the proper nouns (instances)
3. Come up with a list of ontology concepts (concepts)
4. Generate a set of hypothesis phrases using *instance* × *concept* × *pattern*. For example:
 - (a) Pattern (to be described next): e.g. “<instance> is a <concept>”
 - (b) Instance: “U.S.A.”

Methods

1. Extract text from web page
2. Use part of speech (POS) tagging to find the proper nouns (instances)
3. Come up with a list of ontology concepts (concepts)
4. Generate a set of hypothesis phrases using *instance* × *concept* × *pattern*. For example:
 - (a) Pattern (to be described next): e.g. “<instance> is a <concept>”
 - (b) Instance: “U.S.A.”
 - (c) Concept: “country”

Methods

1. Extract text from web page
2. Use part of speech (POS) tagging to find the proper nouns (instances)
3. Come up with a list of ontology concepts (concepts)
4. Generate a set of hypothesis phrases using *instance* × *concept* × *pattern*. For example:
 - (a) Pattern (to be described next): e.g. “<instance> is a <concept>”
 - (b) Instance: “U.S.A.”
 - (c) Concept: “country”
 - (d) Hypothesis phrase: “U.S.A is a country”

Methods

1. Extract text from web page
2. Use part of speech (POS) tagging to find the proper nouns (instances)
3. Come up with a list of ontology concepts (concepts)
4. Generate a set of hypothesis phrases using *instance* \times *concept* \times *pattern*. For example:
 - (a) Pattern (to be described next): e.g. “<instance> is a <concept>”
 - (b) Instance: “U.S.A.”
 - (c) Concept: “country”
 - (d) Hypothesis phrase: “U.S.A is a country”
5. Evaluate set using Google’s hits for each item (Markert et al.)

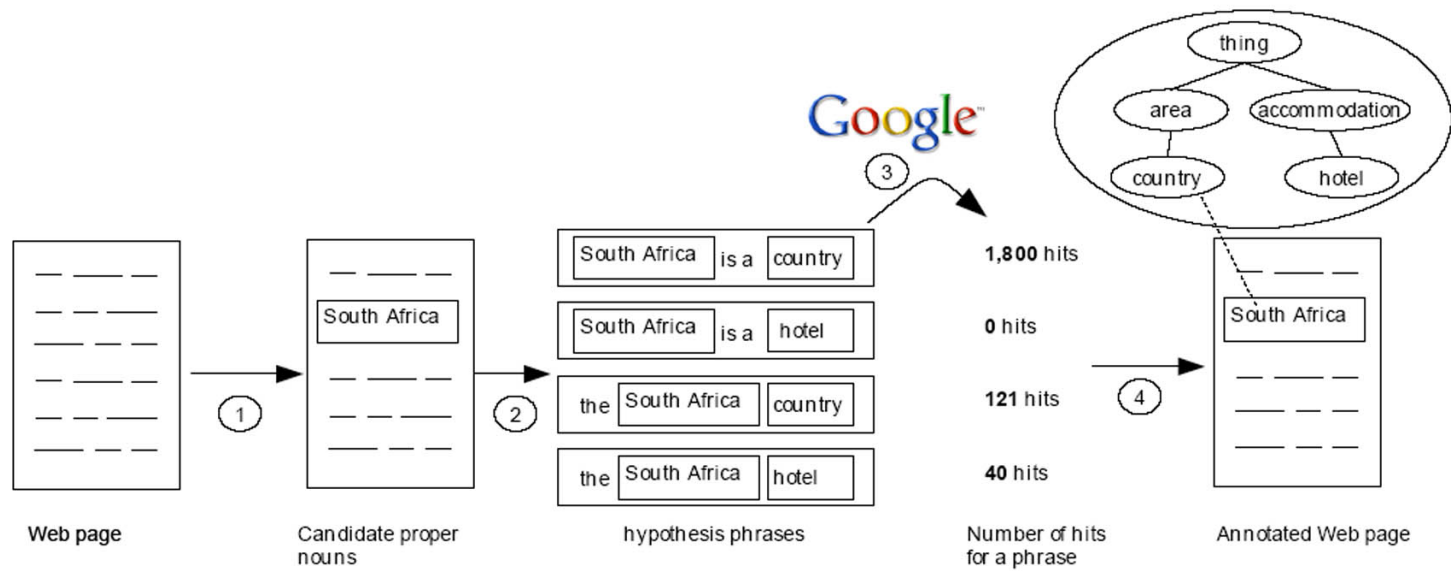
Methods

1. Extract text from web page
2. Use part of speech (POS) tagging to find the proper nouns (instances)
3. Come up with a list of ontology concepts (concepts)
4. Generate a set of hypothesis phrases using *instance* \times *concept* \times *pattern*. For example:
 - (a) Pattern (to be described next): e.g. “<instance> is a <concept>”
 - (b) Instance: “U.S.A.”
 - (c) Concept: “country”
 - (d) Hypothesis phrase: “U.S.A is a country”
5. Evaluate set using Google’s hits for each item (Markert et al.)
6. Use the number of hits to determine the best concept that describes the instance

Patterns

- Hearst
 - <concept>s such as <instance>
 - such <concept>s as <instance>
 - <concept>s, (especially|including) <instance>
 - <instance> (and|or) other <concept>s
- Definites
 - the <instance> <concept>
 - the <concept> <instance>
- Apposition
 - <instance>, a <concept>
- Copula
 - <instance> is a <concept>

Methods



Evaluation

- 30 Texts from www.lonelyplanet.com evaluated by two subjects
- Compare top-n results generated by system with the standards.

- Baseline:

$$count_b(i, c) = \sum_{p \in P} count(i, c, p)$$

- Weighted:

$$count_{\vec{w}}(i, c) = \sum_{p \in p} w_p count(i, c, p)$$

- Use standard precision/recall methods. ($R_{b, \theta}$ = baseline set retrieved with hit cutoff threshold θ , $Standard_y$ = standard set produced by subject y , I = the instance set)

$$Precision = \frac{|\text{correct answers}|}{|\text{total answers}|} = \frac{|R_{b,\theta} \cap Standard_y|}{|R_{b,\theta}|}$$

$$Recall = \frac{|\text{correct answers}|}{|\text{answers in reference standard}|}$$
$$= \frac{|R_{b,\theta} \cap Standard_y|}{|I|}$$

$$F_{1,y} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Results (Before and After Pattern Weighting)

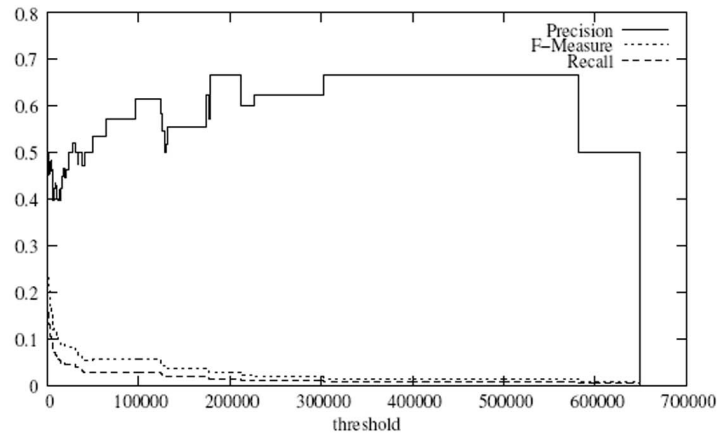


Figure 2: Precision, F-Measure and Accuracy/Recall for $R_{b,\theta}$

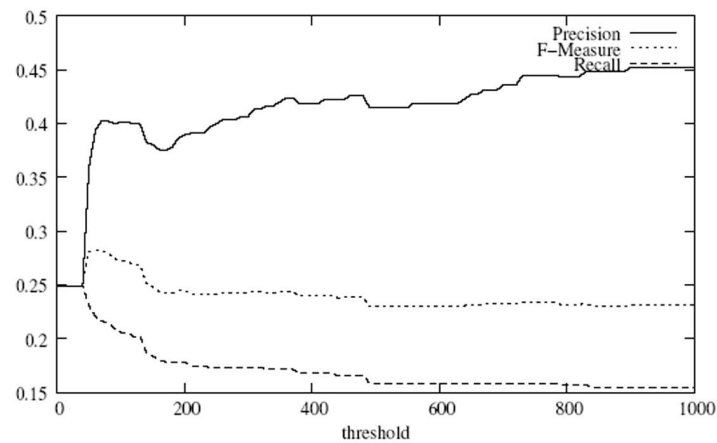
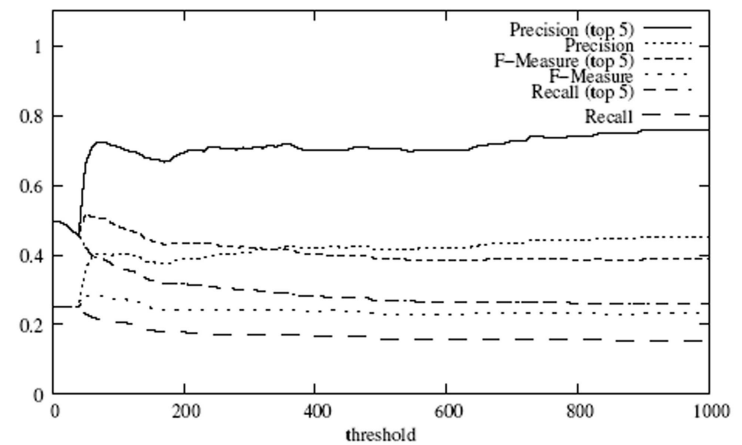


Figure 3: Precision, F-Measure and Accuracy/Recall for $R_{b,\theta}$ zoomed into interval [0..1000]

Contribution / What can be Improved?

- Show the effectiveness of using simple patterns and hits from the search engine in an annotation task (accuracy around 24.9% v.s. 69.09% for human)
- Need to justify the cutoff threshold θ . Should show how the number of hits from Google reflect on the performance of the system
- Accuracy measure misleading for high threshold θ
- In Figure 4, misleading precision for the top 5 results measure
 - Achieve a 30% F_1 improvement with the weighting? Not really.
- Efficiency (277 nouns, 59 concepts, and 10 patterns = 163,430 hits to Google!)

Discussion Questions

- Is this really practical in the context of semantic web?
- Does this scale?
- Who would use this system?
- Is there a better way to leverage the web as a knowledge base for annotation?