Dublin Core Conference, Berlin, September 22-25, 2008

Semantic Relation Extraction from Socially-Generated Tags: A Methodology for Metadata Generation

Miao Chen, Xiaozhong Liu, Jian Qin
School of Information Studies
Syracuse University
Syracuse, NY, USA

Agenda

Semantic relation extraction

- Background
- Methodology and solution design
- Experiment
- Result and analysis
- Conclusion

The Social Spaces

Semantics defined

Linguistics:
Study of meaning at the level of words, phrases, sentences, and even larger units of discourse.

Computer science:
An application of mathematical logic; the meaning of programs or functions

Psychology:
Relationship between memory and meaning. Semantic memory is memory for meaning.

Philosophy:
Study of symbolic logic: meaning and truth, meaning and thought, and the relation between signs and what they mean.

Sources: The Columbia Encyclopedia, Sixth Edition. 2008. http://www.encyclopedia.com/doc/1E1-semantic.html; Wikipedia. http://en.wikipedia.org/wiki/Semantics

Semantics in social spaces

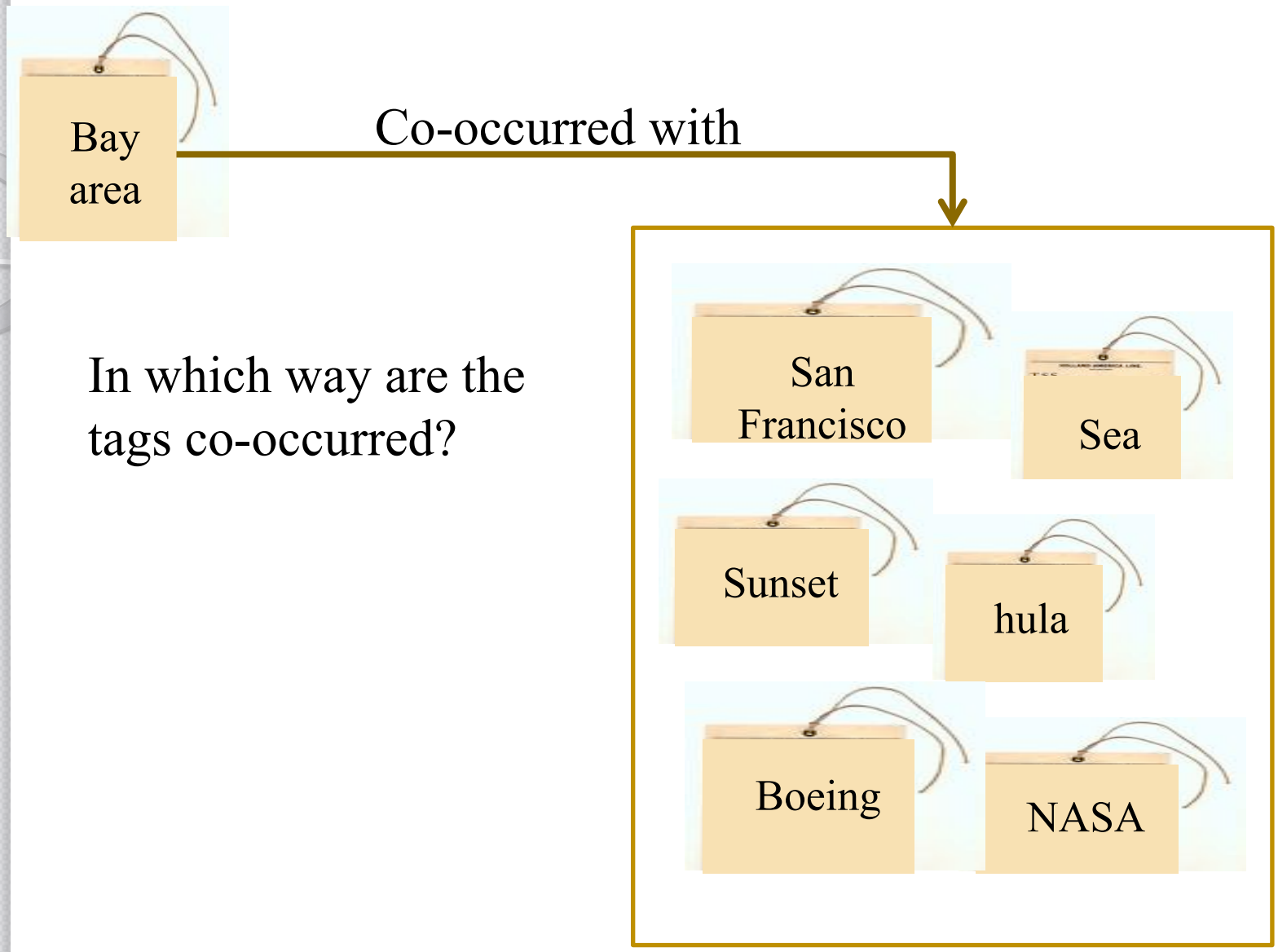
- Study of meanings and relationships between expressions in social spaces
  - Tags as the agents for semantics in social spaces can be studied
    - as words (linguistics)
    - as programmatic symbols (computer science)
    - as memory for meaning (psychology)
    - as symbolic logic (philosophy)
    - as metadata (library and information science)

Characteristics of semantics in social spaces

- Community-based
  - Focus on a group of relevant topics/concepts
  - Sharing common interest among members
  - Dynamic, evolving scopes of semantics
- Randomly generated by independent users
  - Personal view and perception of the world
  - Varying quality of information
  - High volume of information noise

Extracting semantic relations in social spaces

- Tag mining can generate valuable subject metadata created by users
- Concepts and relationships extracted can be used to enhance subject metadata



In which way are the tags co-occurred?

Challenges in tag mining

- Lack of semantic relations between tags
- Lack of context under which tags are assigned
- Massive tag data sources that require automatic processing

Strategies to the challenges

- Extracting semantic relations from text data can use
  - Statistical approach

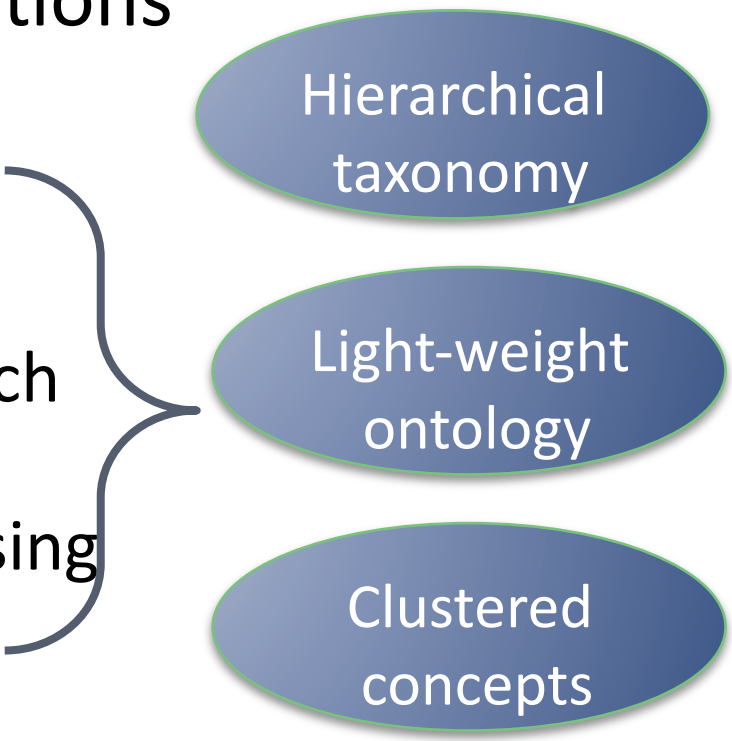
(Heymann & Garcia-Molina, 2006; Schmitz, 2006)

  - Machine learning approach

(Bunescu & Mooney, 2007; Culotta & Sorensen, 2004)

  - Natural Language Processing (NLP) approach

(Bunescu & Mooney, 2007; Zelenko et al., 2003)

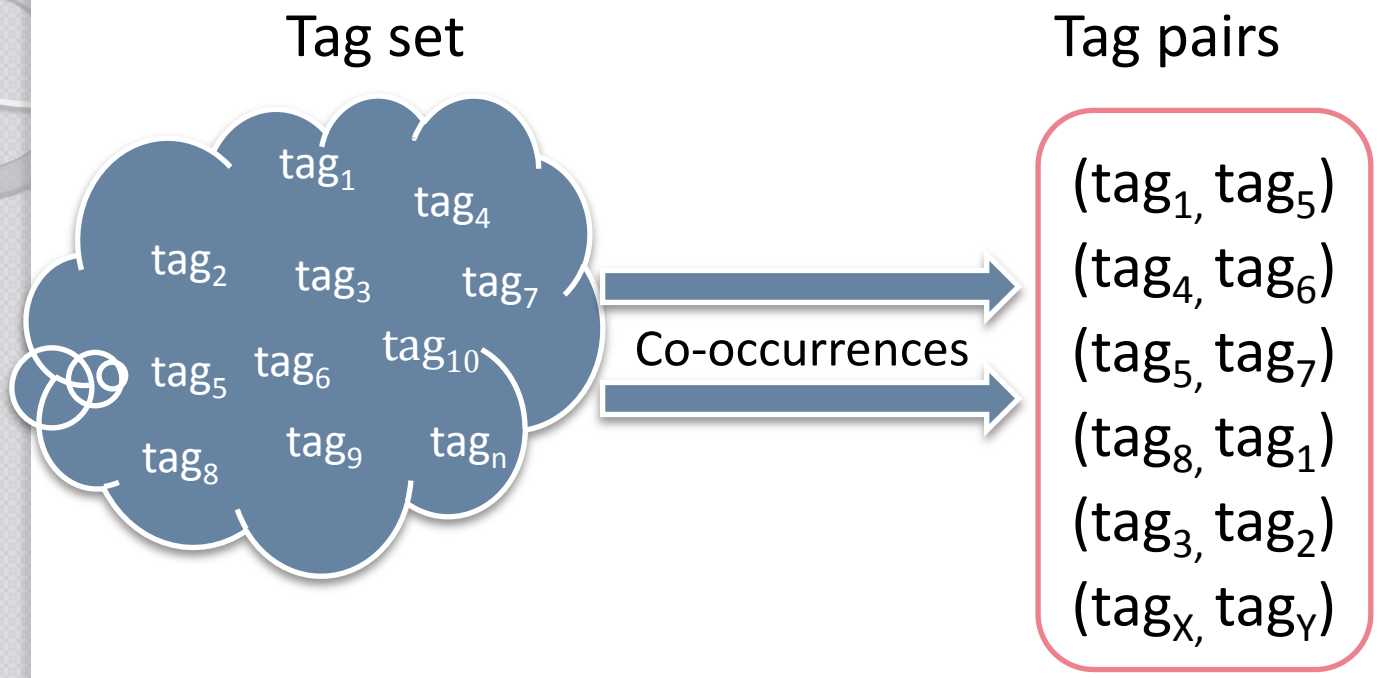


But in tag mining, the *context* is missing from the scene…

Introducing context in the process

- An approach that combines
  - NLP techniques
  - Machine learning techniques
  - *Context information* obtained from using Google

Methodology: Identifying tag pairs

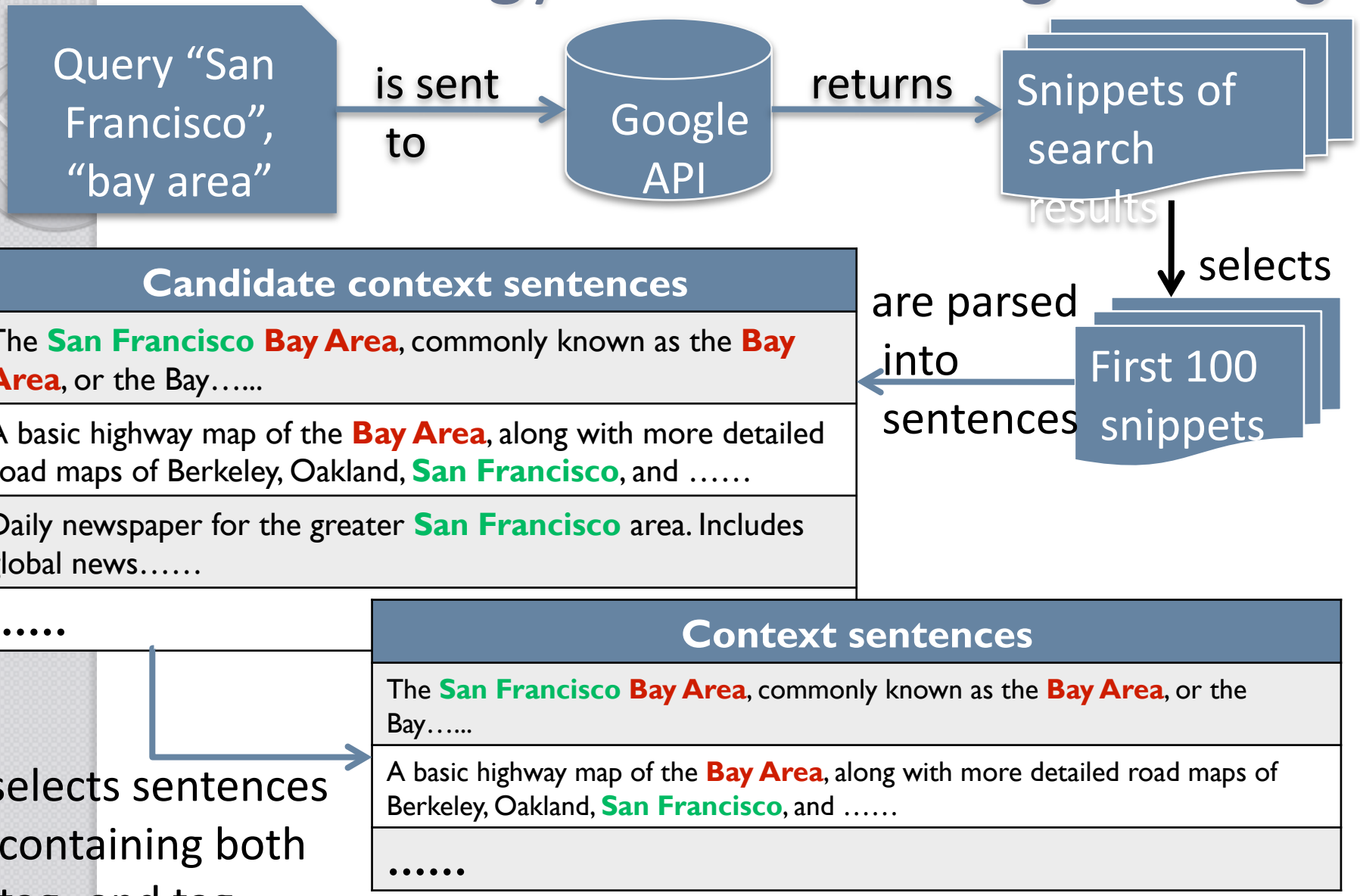


High co-occurrences → High possibility of relation between tags

tag_X = San Francisco
tag_Y = Bay area

Relation (tag_X, tag_Y) = ?

Methodology: Context info gathering



Methodology: Machine learning

A parsed context sentence

The/DT largest/JJS city/NN in/IN the/DT Sonoran/NNP Desert/NNP is/VBZ

Phoenix/NNP , Arizona/NNP , U.S.A./NNP

Selected features

1) Bag of words features

i.e. [text window, size n, Porter stem] largest, citi, in, sonoran, desert

2) Position features

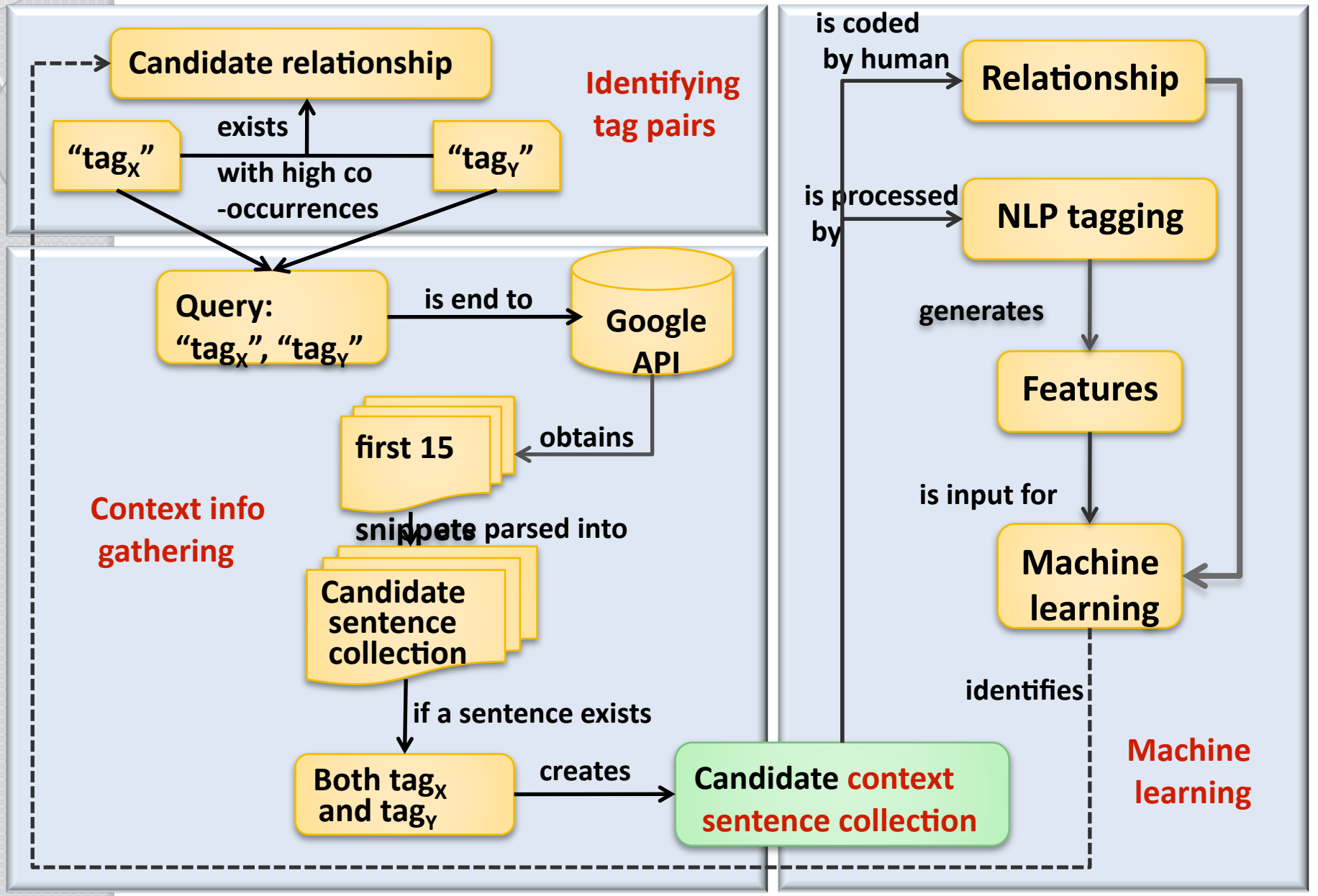
i.e. word between tagX and tagY, before tagX Position, after tagY Position

3) Semantic features

i.e. Lexical: part-of-speech JJS, NN, IN DT

Syntactic: NP [DT NNP NNP; the sonoran desert]

Methodology: Relation extraction



Data

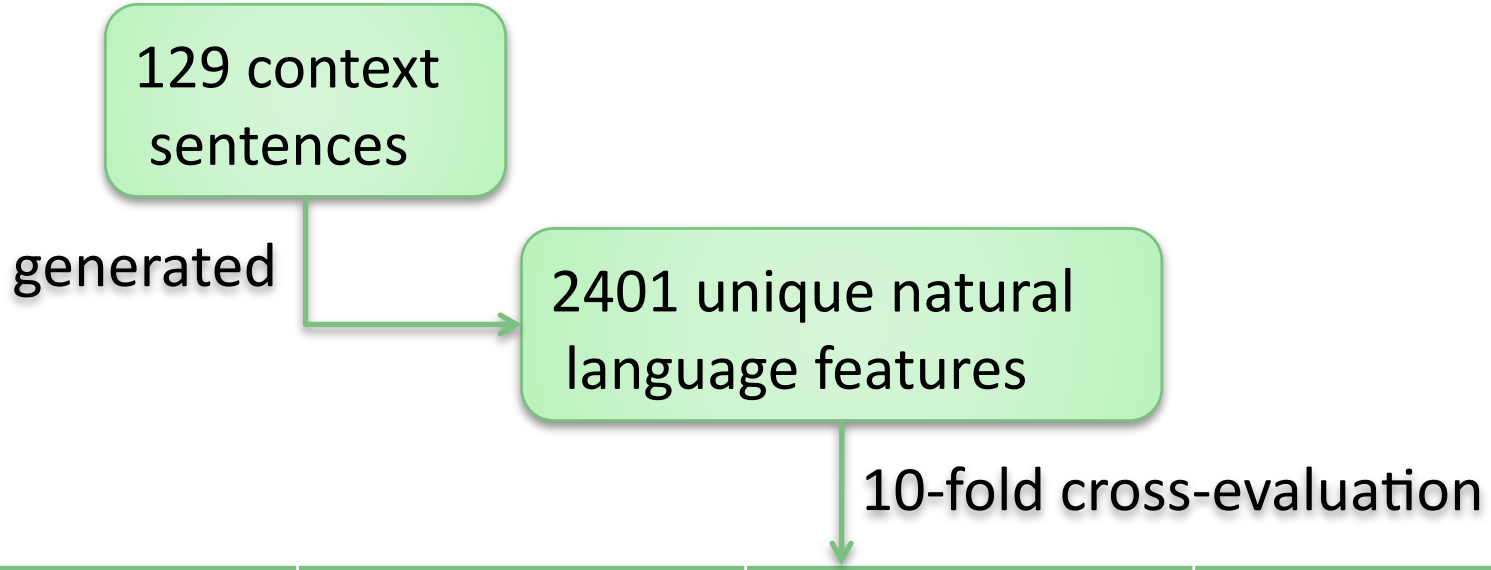
- Sample size: 28,737 photos about "landscape" from Flickr
- Total number of tags: 289,216 (21,443 unique)
- First 3,000 tag pairs with highest mutual information score were selected for relation extraction.

Experiment: result

Relations between tag_X and tag_Y are returned as experiment results. In this preliminary study, three types of relations are identified, as shown below.

tag_X	relation	tag_Y	Context sentence
2-deoxy-d-glucose	induces	effect	Effect of 2-deoxy-D-glucose on cell fusion induced by Newcastle disease and herpes simplex viruses.
Action	is-induced-by	anticonvulsant	Pharmacokinetic modeling of the anticonvulsant action of phenobarbital in rats. J Dingemanse, JB van Bree and M Danhof.
Acadia	is-located-in	maine	Use this vacation and travel guide to the Downeast and Acadia region of Maine to plan your vacation, business trip or just for fun.

Evaluation



	induces	is-induced by	is-located-in
induces	90	2	1
is-induced by	10	12	3
is-located-in	1	4	6
Correctly classified instances: 108 (83.72%)			
Incorrectly classified instances: 21 (16.28%)			

Discussion

- A hybrid methodology for extracting semantic relations between socially-generated tags
- CONTEXT information can facilitate tag disambiguation
- Promising applications in
  - Automatic subject metadata generation
  - Automatic subject vocabulary mapping
  - Development and enrichment of controlled vocabularies

Future Work

- Use a larger training data set and test more types of relations
- Experiment with extracted relations to examine its performance
- Integrate more resources to acquire richer context information
- Apply the methodology to other social tagging resources

References

- Heymann, Paul, and Hector Garcia-Molina. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Department of Computer Science, Stanford University. Retrieved, April 13, 2008, from http://labs.rightnow.com/colloquium/papers/tag_hier_mining.pdf
- Schmitz, Patrick. (2006). Inducing Ontology from Flickr Tags. Collaborative Web Tagging Workshop at WWW 2006, Edinburgh, UK. Retrieved, April 13, 2008, from http://www.topixa.com/www2006/22.pdf
- Bunescu, Razvan. C., and Raymond J. Mooney. (2007). Extracting relations from text from word sequences to dependency paths. In A. Kao et al. (Ed.): Text Mining and Natural Language Processing (pp. 29-44). London: Springer.
- Culotta, Aron, and Jeffrey Sorensen. (2004). Dependency tree kernels for relation extraction. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Retrieved, April 13, 2008, from http://acl.ldc.upenn.edu/P/P04/P04-1054.pdf
- Zelenko, Dmitry, Chinatsu Aone, and Anthony Richardella. (2003). Kernel methods for relation extraction. Journal of Machine Learning Research, 3, 1083-1106.
- www.flickr.com