UIMA Workshop, GLDV, Tübingen, 09.04.2007

Iterative Learning of Relation Patterns for Market Analysis with UIMA

Sebastian Blohm, Jürgen Umbrich,
Philipp Cimiano, York Sure

Universität Karlsruhe (TH), Institut AIFB
blohm@aifb.uni-karlsruhe.de

AIFB

Motivation

- A **lot of facts** on the Web are not available in structured form. But we would like to have them structured.
- The **Web is big**. For an individual user task, linear-time processing is prohibitive.
- We need to be able to **derive information on demand** and thereby **take advantage of previous annotations**.
- **Classical Web search** indices allow fast access, but only for **pure text**.
- **Structural queries** also allow this but **require knowledge on the structure of the content**.
- We therefore want to **learn structured queries** that combine classical and semantic indices.

AIFB

Project context of this work

AIFB

Outline

- Iterative Induction of Patterns
- Going for structured queries
- How to make structure learnable
- Status of work

AIFB

Iterative Pattern Induction

- Early text mining information extractors heavily relied on manually defined extraction patterns [Hearst92]. Automatic generation of patterns:
  - Reduces work
  - Increases flexibility
  - Allows population of ontologies with many different relations.
- Our approach:
  - Input: Few instances of a relation

  - Process: Use Web search to identify how relation instances are typically mentioned.

  - Output: Patterns that allow extracting many instances through web search.

AIFB

Learning Patterns from Occurrences

All possible merges of patterns are considered. Example merge:

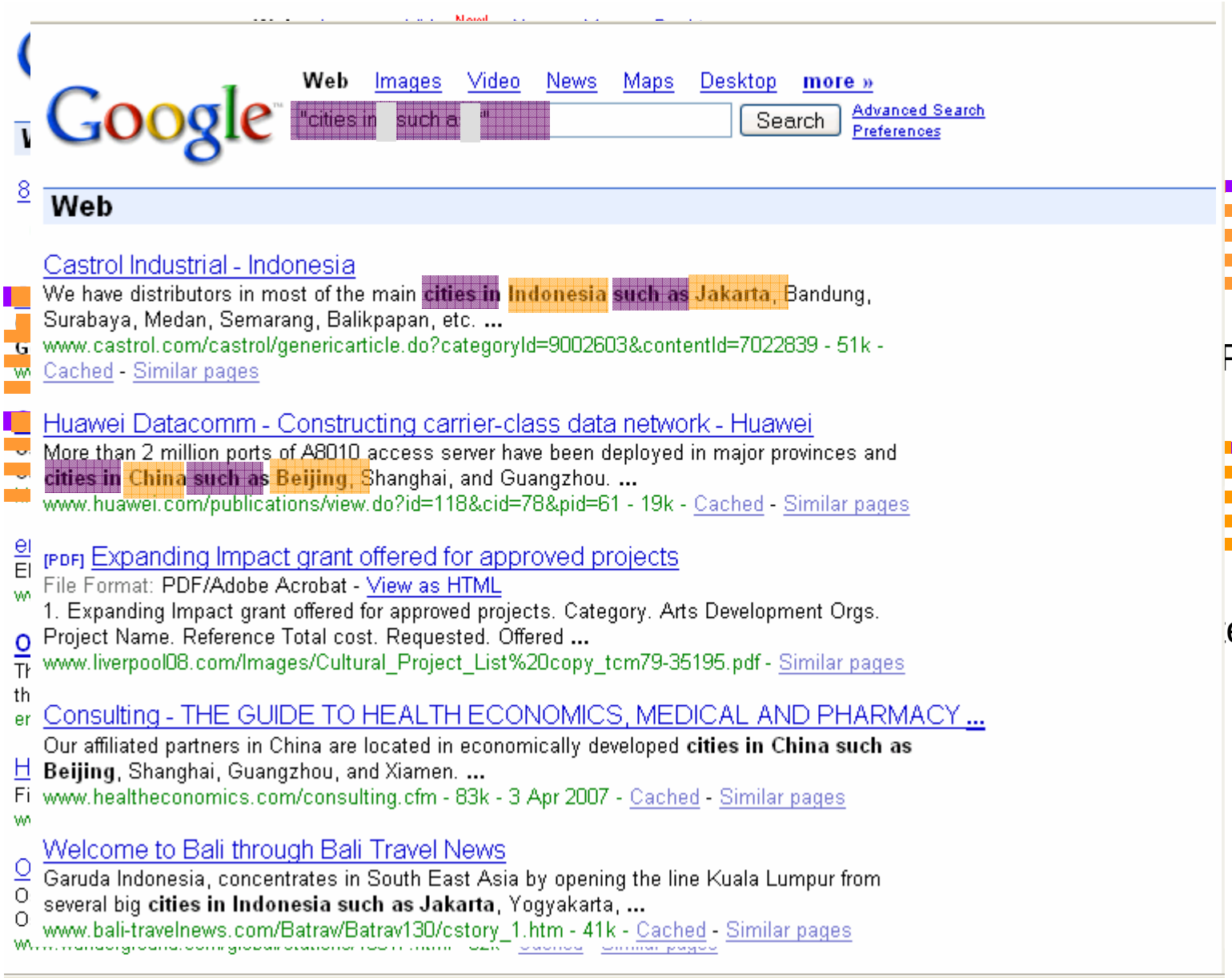
The happiest people in Germany live in Osnabrück .
The richest people in America live in Hollywood.
The * people in * live in * .

Related Work

- Static Patterns [Hearst 1992]
- Bootstrapped Learning on search index [Brin 1998]
- Wrapper Induction [Kushmerick 2000]
- Large Scale Systems [Etzioni et al., 2005]

AIFB

The PRONTO system



Patterns

erns

AIFB

Design Choices

AIFB

Structure of Patterns

- Lists of words (cleaned)
- Only occurrences with a max argument distance of 4 are considered.
- Window of processing: 2 words before the first and after the last argument.
- Punctuation is kept (punctuation chars are distinct words)
- Capitalization is checked for.

Nature of queries

Tuples: just full text of the arguments

Patterns: quote, use * wildcard, remove surrounding wildcards

```
"flights to  * ,  * from northeast“
```

Going for more complex patterns

Clearly, processing **would benefit from**

- Gazetteers
- Shallow linguistic processing
- Other UIMA annotators

This **leads to**:

- better extraction performance
- general patterns that can be used for large scale annotation (sub-linear performance)

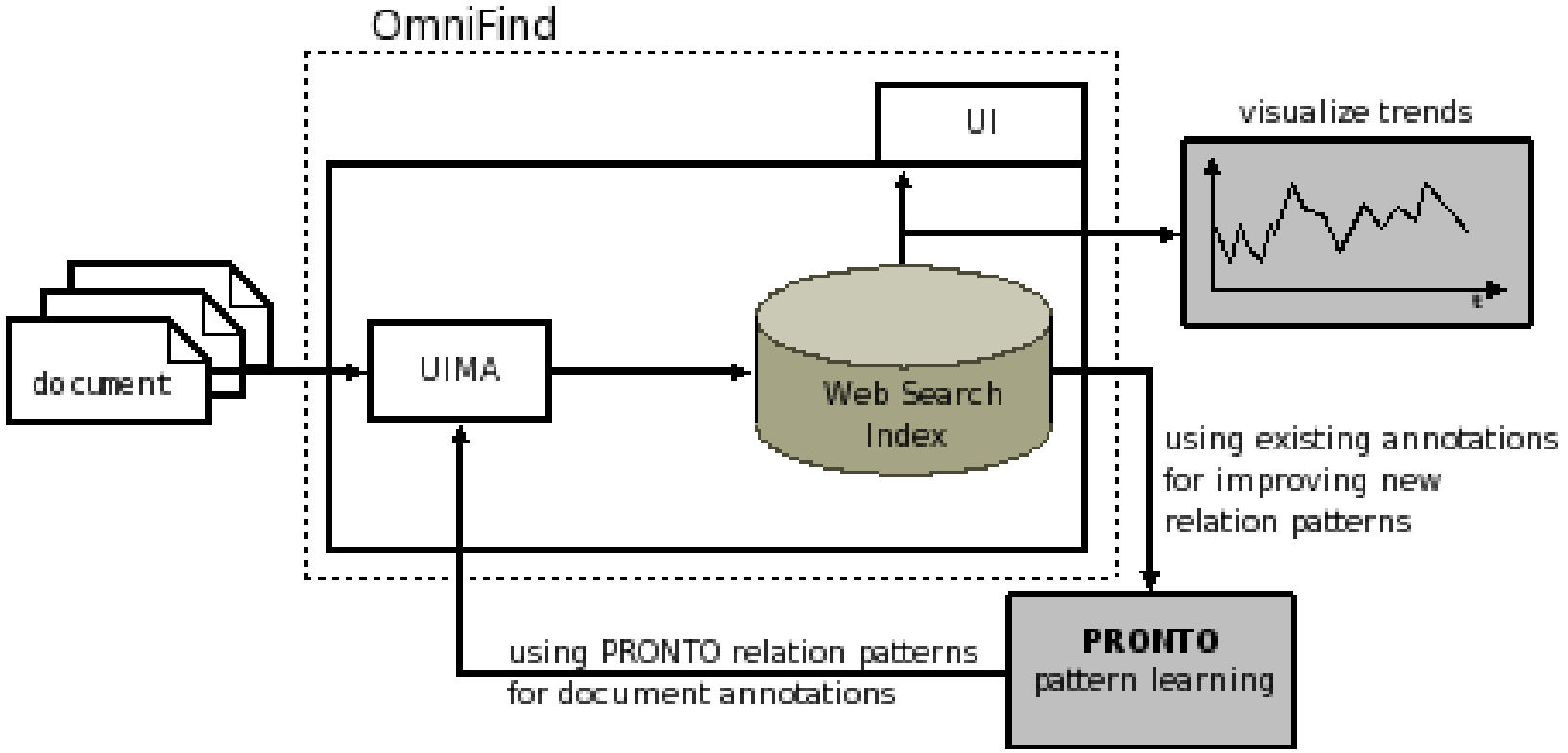
... but it would need to **learn, how to employ the annotations**.

This means, we need to **formalize text and annotations** in a way that allows:

- Structural querying
- Abstraction for learning

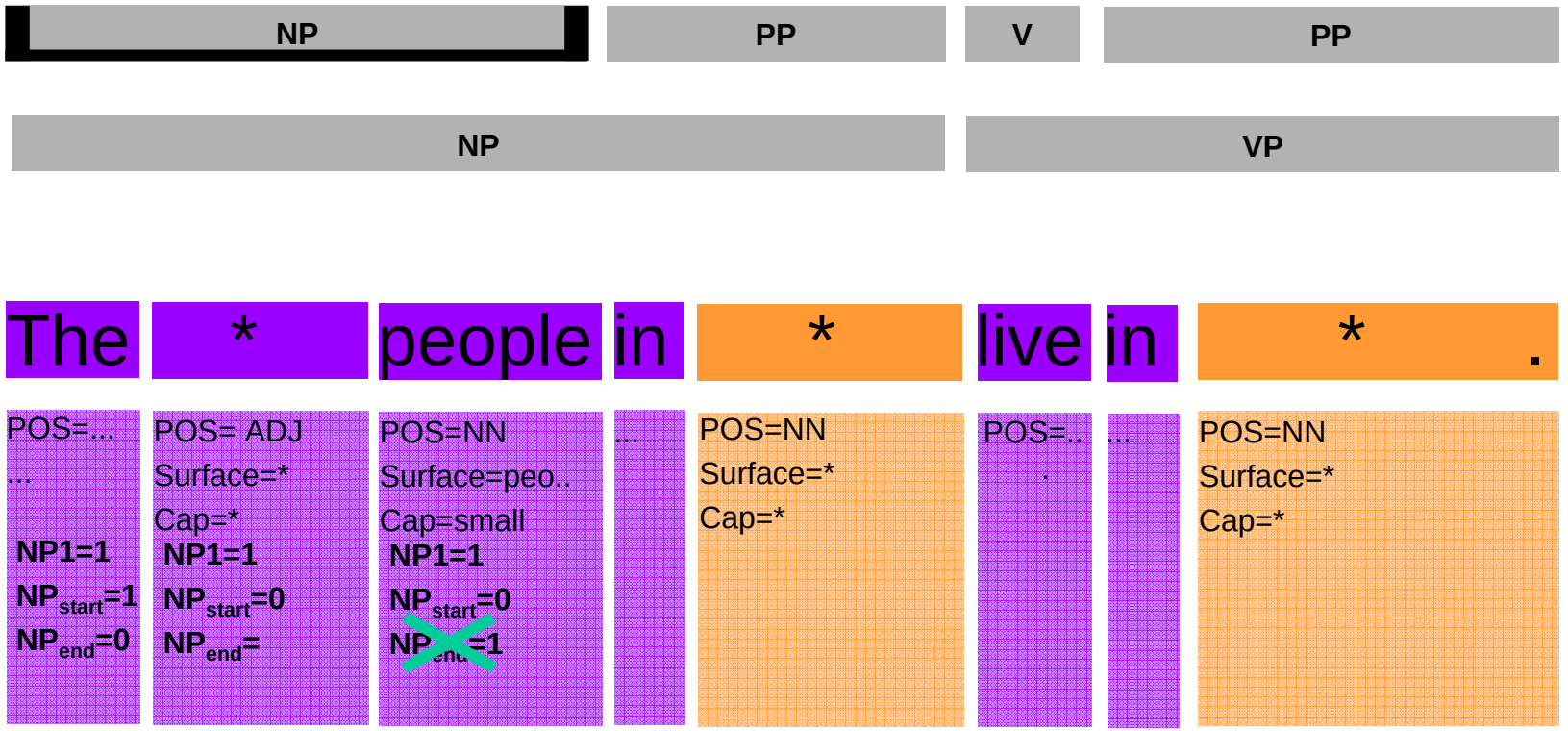
AIFB

Where UIMA comes into play



AIFB

Representing Annotations in Patterns (sub-optimal)



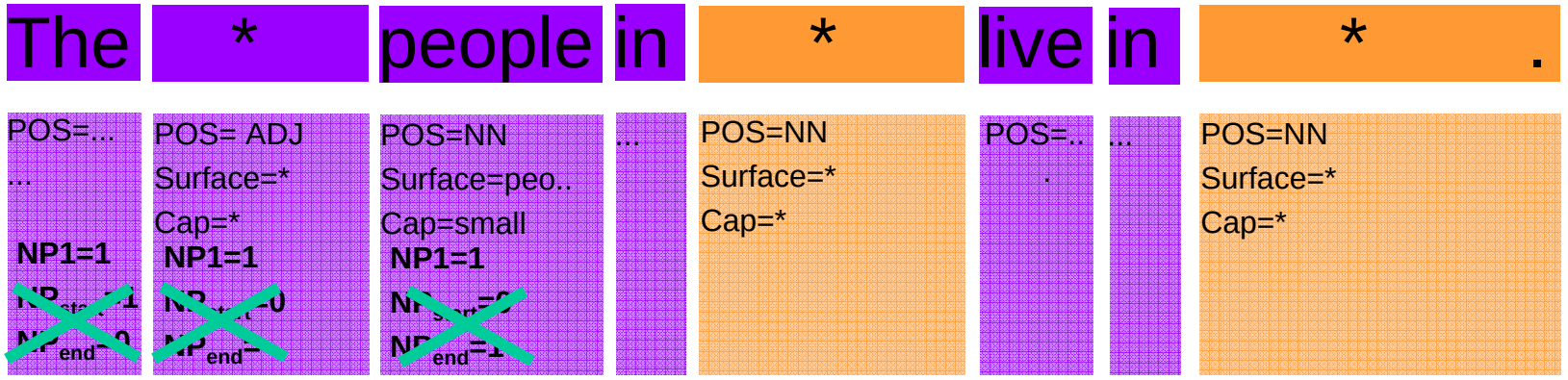
- For the learning phase, patterns are represented as feature vectors for each token.
- UIMA Annotations indicate spans of text.
- Translation: Represent beginning, end and arbitrary position
- Learning consists of eliminating too specific features

AIFB

Querying for complex patterns

Key points:

- Combine **textual matches** with **structual matches**
- Enforce **order** but not everywhere
- Make **annotations** as "**atomic**“ as possible to allow abstraction along many dimensions.
- Is annotation **overload** an issue?




```
<S>
  <NP>“The“<token POS="ADJ“/>“people in“</NP>
  <#token POS="NN"/>“live in“<#token POS="NN"/>
</S>
```

Status of Work

PRONTO System

- Ready for Web extraction with pure text patterns [AAAI 07]
- Exposed Plug-In API: almost there

UIMA Integration

- Annotators to identify objects of various classes: done
- Integration with OmniFind: 80% done
- Matching procedures: ongoing

Future Plans

- Visualization for market analysis
- Smarter pattern learning
- Any ideas?

AIFB

Thank you for your attention

Sebastian Blohm, Jürgen Umbrich,
Philipp Cimiano, York Sure

Universität Karlsruhe (TH), Institut AIFB
blohm@aifb.uni-karlsruhe.de

AIFB

References

[Hearst92] M. A. Hearst, \Automatic acquisition of hyponyms from large text corpora," in Proceedings of the 14th conference on Computational linguistics. Morristown, NJ,USA: Association for Computational Linguistics, 1992, pp. 539-545.

[DIPRE98] S. Brin, \Extracting patterns and relations from the world wide web," in *WebDB Workshop at 6th International Conference on Extending Database Technology,EDBT'98*, 1998.

[KnowItAll05] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland,D. S. Weld, and A. Yates, \Unsupervised named-entity extraction from the web: an experimental study," Artif. Intell., vol. 165, no. 1,

[Snowball00] E. Agichtein and L. Gravano, \Snowball: extracting relations from large plain-text collections," in *DL '00: Proceedings of the fth ACM conference on Digital libraries*. New York, NY, USA: ACM Press, 2000

[Espresso06] M. Pennacchiotti and P. Pantel, \A bootstrapping algorithm for automatically harvesting semantic relations," in Proceedings of Inference in Computational Semantics (ICoS-06), Buxton, England.

[CIA01] http://www.daml.org/2001/12/factbook/

[AAAI07] S. Blohm, P. Cimiano and Egon Stemle: "Harvesting Relations from the Web – Quantifying the Impact of Filtering Functions“. In Proceedings of the AAAI 2007. Vancouver, Canada. (to appear)

[Etzioni et al., 2005] Oren Etzioni, Michael J. Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, Alexander Yates: Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence* 165(1): 91-134 (2005)