Text Mining for Intellectual Property

C.H.A. Koster and N. Oostdijk,
Radboud University, Nijmegen, The Netherlands
H. Berger,
Matrixware, Vienna, Austria

1. Project goals

The goals of the TM4IP project are:

- to develop a professional search engine based on deep linguistic techniques (PHASAR)
- to develop an accurate parser for complicated technical english texts (AEGIR)
- to combine these into a Text Mining system for Intellectual Property search.

The project is a collaboration between the departments of Linguistics and Computer Science of the University of Nijmegen and Matrixware.

2. PHASAR

The PHASAR prototype was developed in the Dutch NBIC programme as a literature mining system, providing an alternative form of access to the Medline collection of medical abstracts.

- PHASAR uses Dependency Triples as search terms instead of keywords, achieving a much greater precision
- it provides its user with feedback from the document index, supporting explorative search and query specialization
- it provides quantified thesauri to support query generalization

In this way it gives the searcher full control over precision and recall in a transparent way.

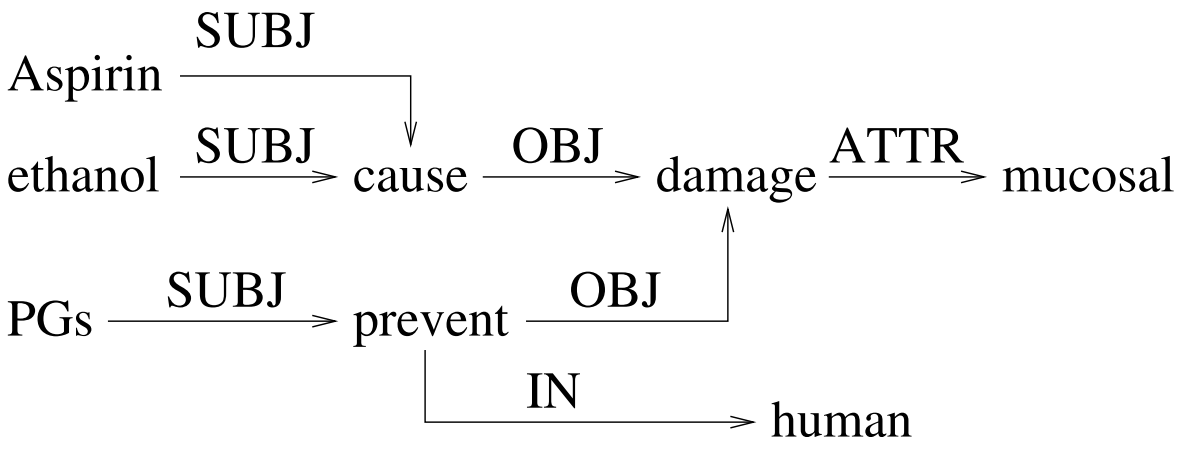
2.1 Dependency Trees

By a *dependency tree* we mean a graph (a tree with possibly additional confluent arcs) whose nodes are marked with words and whose arcs are marked with directed syntactic relations.

As an example, the sentence

In humans, PGs prevent the mucosal damage caused by aspirin and ethanol

results in the following dependency "tree" (after lemmatization and transforming one of the sentences from passive to active):



2.2 Dependency Triples

A *dependency triple* is a triple [word,relation,word] obtained by unnesting from a dependency tree.

relation	*example*
subject relation	`[PG,SUBJ,prevent]`
object relation	`[prevent,OBJ,damage]`
predicate relation	`[Aspirin,PRED,painkiller]`
attribute relation	`[damage,ATTR,mucosal]`
attribute relation	`[consumption,ATTR,ethanol]`
prepos relation	`[consumption,of,ethanol]`
prepos relation	`[rely,on,measurement]`
prepos relation	`[effective,against,bleeding]`
modifier relation	`[increase,MOD,not]`

In PHASAR both the documents and the queries are represented by dependency triples, invisible to the user.

3. AEGIR

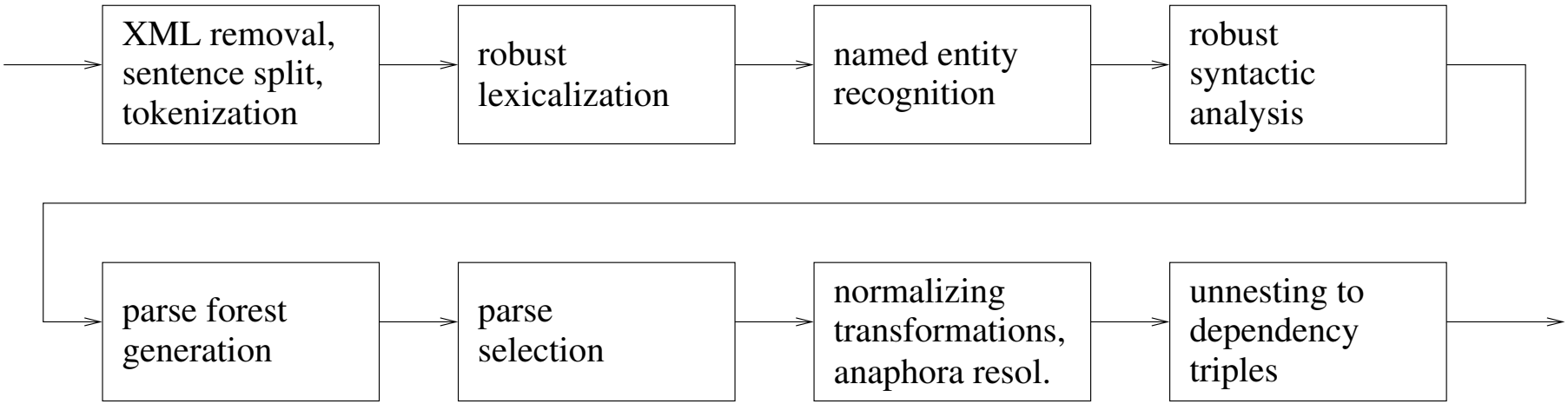
AEGIR ("Accurate English Grammar for IR") is a dependency parser for English.
It is a hybrid parser, consisting of

- a rule-based grammar and its associated lexica, representing (in Chomskian terms), the "language competence"
- a database of reliable Dependency Triples, representative for the application and gathered from various sources, representing language use.

The AEGIR grammar is written in the AGFL formalism, producing fast parsers. The database of triples provides high accuracy.

3.1 Traditional parsers

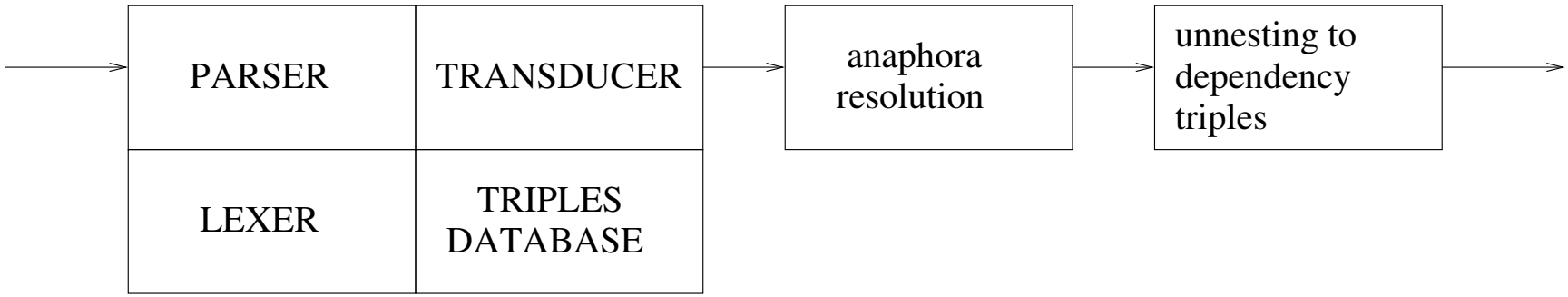
- Patent texts are among the hardest human-readable documents to analyse syntactically, requiring hybrid parsers trained on domain material.
- Using traditional techniques, the parser would be a long pipeline:



- strictly sequential execution
- a bad choice made in an earlier stage can never be corrected.

3.2 AGFL parsers

- AGFL = weighted attribute grammar with set-valued features
- the AGFL system generates Top-Down chart parsers with the Best-Only heuristic
- combining different sources of probabilities and preferences



- fast parsing: 2400 words/second on a PC.

4 Text Mining?

- Text Mining = Search + Analysis
- State-of-the-art: document co-occurrence of terms
  - works well for abstracts only
- PHASAR also offers sentence co-occurrence of terms
  - works on full-text documents as well
  - patent applications, journal articles and dissertations
  - even the whole internet
- PHASAR supports analysis
  - classification techniques for document selection and presentation
  - search within search
  - interactive construction of re-useable search profiles
  - aggregating information from different documents.

5 Intellectual Property?

Key observation: patent searchers prefer Boolean search over ranked search

- need for completeness
- need for transparency and accountability
- desire for full control over precision and recall
- willingness to invest work to achieve these

PHASAR offers

- exact match with full transparency
- explicit mechanisms for control over precision and recall
- qualitative and quantitative feedback from index and thesaurus

PHASAR is what patent searchers need.