Natural Language Watermarking: Challenges in Building a Practical System

Mercan Topkara
Mikhail J. Atallah

Department of Computer Sciences,
Purdue University

Giuseppe Riccardi

Department of Information and Communication Technology,
University of Trento

Dilek Hakkani-Tür

AT&T Labs-Research

Problem and Key Idea

Performing natural language watermarking that is quantifiably high quality while being easy to use

A natural language watermarking system that

- is based on linguistic transformations
- follows *language engineering* principles
- is tested against well-known benchmarks

Outline

- **Natural Language Watermarking**
  - Applications
  - Challenges
- **Watermarking at Sentence Level**
- **Experimental Setup**
- **Results and Quantitative Evaluation**
- **Conclusions and Future Work**

Applications

- **Meta-data binding**
- **Defense against phishing**
  - Controlling distribution and reuse of intellectual property
  - Proving or denying ownership on a document
- **Enforcing access control policies**
  - Multi-party private communications
  - Digital libraries
- **Content protection**
  - On-line news channels, on-line stores etc.
- **Text auditing, tamper-proofing, traitor tracing**

Challenges

- **Automatic semantic text analysis and evaluation**
  - Modeling user perception
- **Low bandwidth**
- **Combinatorial syntax and semantics sentences**
- **Achieving robustness**
  - Preserving the characteristics of cover

Challenges

- **Preserving characteristics of cover text**
- **Evaluating characteristics of cover text**
  - Meaning

    "I saw the woman with a telescope." (Noun phrase attachment)
  - Fluency
    - Requires higher level analysis of full text
  - Grammaticality

    "I can can the can." (Part of speech)
  - Style

Previous Work in NL Watermarking

- **Using Syntactic Transformations** (Atallah et. al, 2001)
  - Watermark is embedded into the binary encoding of syntactic trees
  - Applied Transformations
    - Adjunct movement, Clefting, Passivization - Activization
- **Using Semantic Transformations** (Atallah et al., 2002)
  - Applied Transformations
    - Grafting, pruning and substitution of the concepts in the semantic tree

The EU ministers will tax aviation fuel as a way of curbing the environmental impact of air travel.

```
author-event-1--|--author--unknown
                |--theme--levy-tax-1--|--agent--set-4--|--member-type--geopolitical-entity
                                      |                |--cardinality--unknown
                                      |                |--members--(set| "EU nations")
                                      |--theme--kerosene-1
                                      |--purpose--regulate-1--|--agent--unknown-1
                                                              |--theme--effect-1--|--caused-by--flight
```

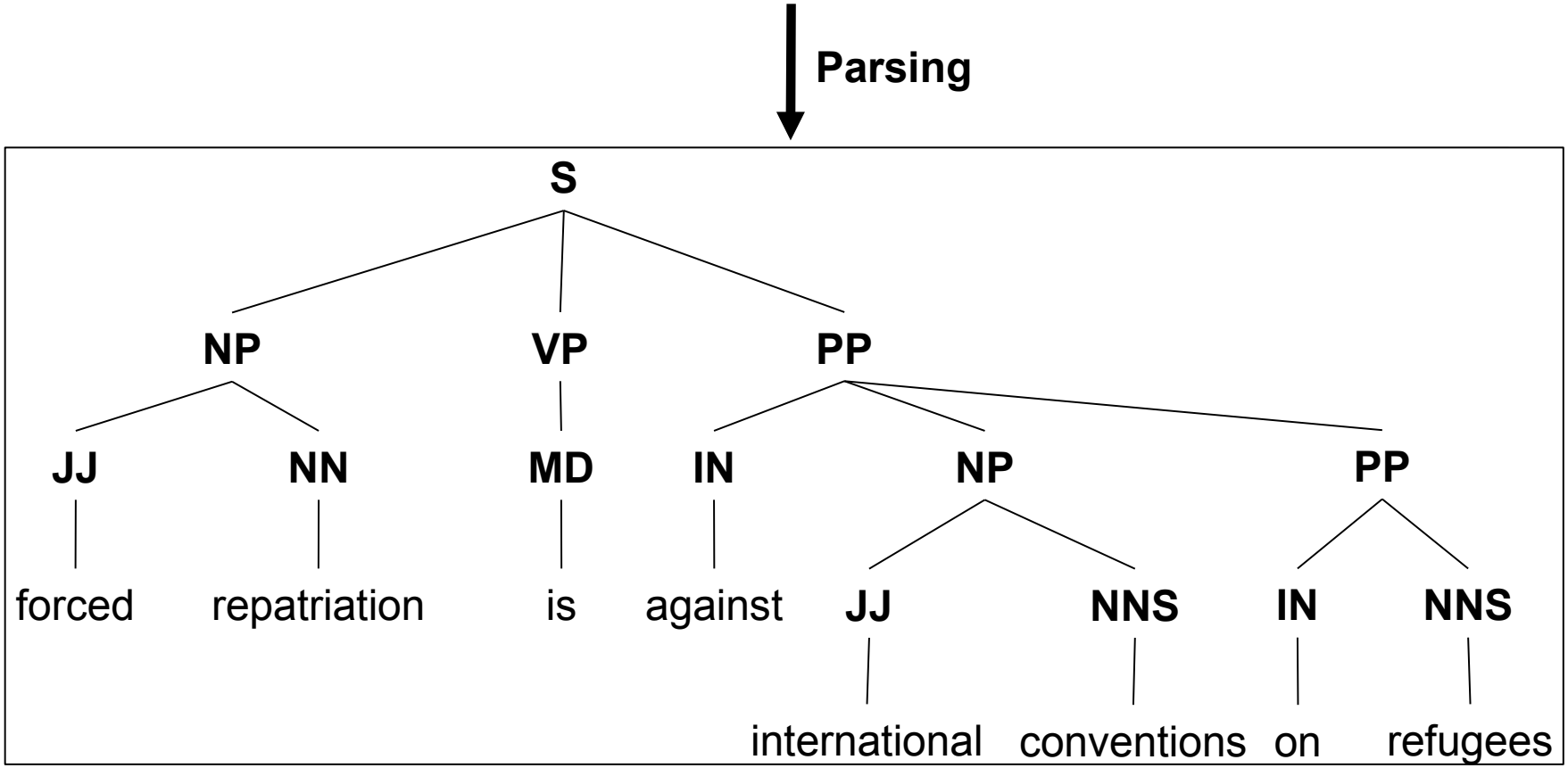
Watermarking at the Sentence Level

- **Linguistic transformations are defined at sentence level**
- **Easier to preserve characteristics of the full text**
- **Electronic Data Resources**
  - Reuters Corpus, WordNet, VerbNet
- **Natural Language Parsers at Sentence Level**
  - XTAG, Charniak, Collins, Stanford, Links…
- **Natural Language Generators at Sentence Level**
  - Realpro

Watermarking at Sentence Level

- **Natural Language Parsing**
  - Syntactic Parse Tree

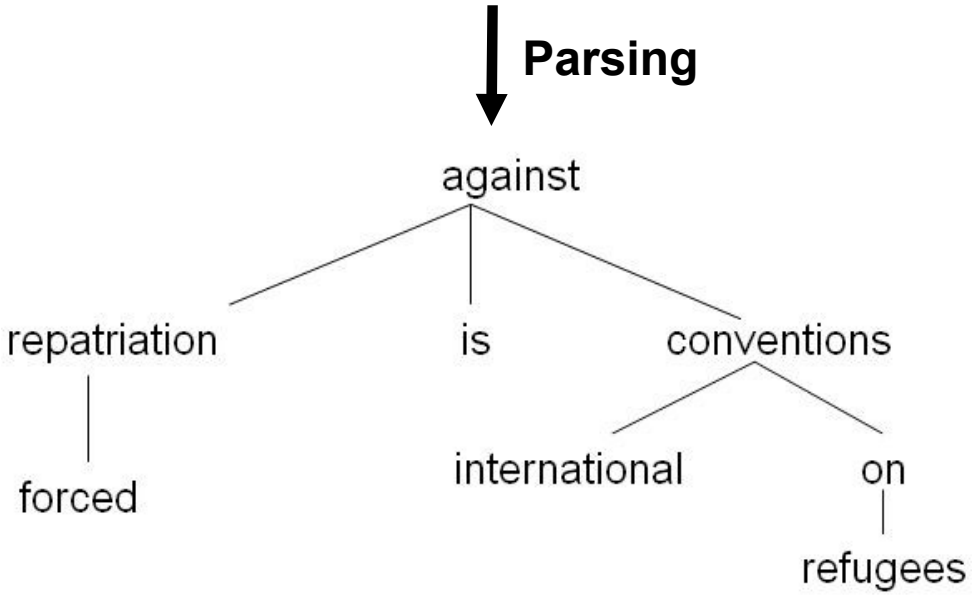
Forced repatriation is against international conventions on refugees.



Watermarking at Sentence Level

- **Natural Language Parsing**
  - Dependency Tree Generation

Forced repatriation is against international conventions on refugees.



Dependency Tree Structure

Watermarking at Sentence Level

- **Natural Language Generation**

Deep Syntactic Structure (DSyntS)

```
      against [ class:preposition ]
(  I  repatriation [ class:common_noun article:no-art number:sg ]
        ( ATTR  forced [ class:adjective ] )
  II  convention [ class:common_noun article:no-art number:pl ]
        ( ATTR  international [ class:common_noun article:no-art number:sg ]
            II  on [ class:preposition ]
              ( II  refugee [ class:common_noun article:no-art number:pl ] ) )
 III  be [ class:verb number:sg person:3rd case:nom tense:pres aspect:simple ] )
```

Sentence realization

Forced repatriation is against international conventions on refugees.

PURDUE UNIVERSITY

Sentence Level Linguistic Transformations

- **Synonym Substitution**
- **Syntactic Transformations**
  - Passivization, topicalization, clefting, preposing, there-construction, fronting

    "He is in critical condition, doctors said."
  - Verb Alternations (Levin Verb Classes):
    - 193 verb classes, covering 3100 verbs

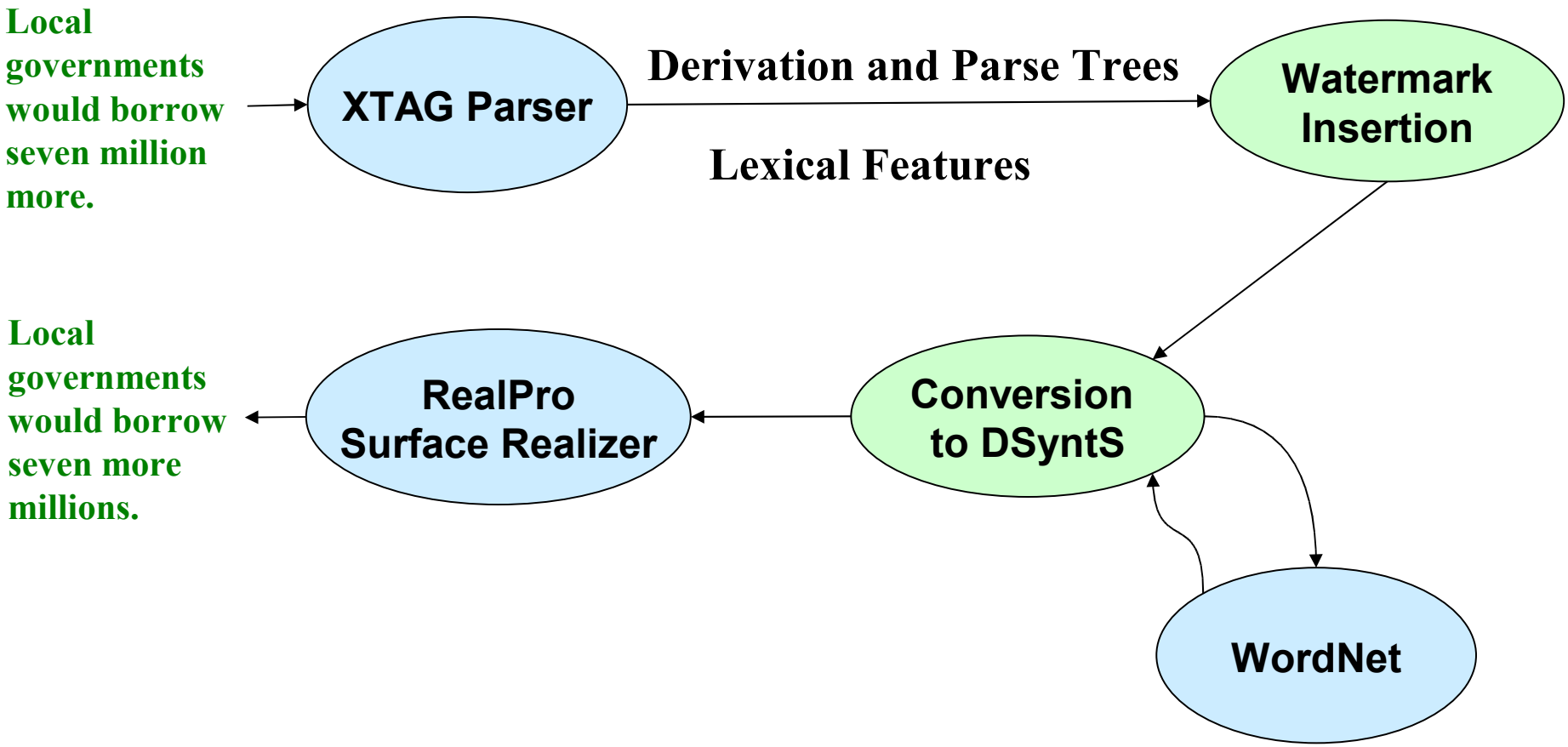
"Mary sold Amy a car." ⟺ "Mary sold a car to Amy."

- **Semantic Transformations**

Experimental Setup

- **Data**
  - Reuters newswire reports from 24 August 1996, 20 October 1996, and 19 August 1997
  - Filtered out the sentences that can be parsed in less than 30 seconds
  - 683 sentences
- **Tools**
  - XTAG Parser
  - Charniak Parser
  - Lextract
    - Conversion of phrase structures to dependency structure
  - Realpro
    - Sentence level surface realization

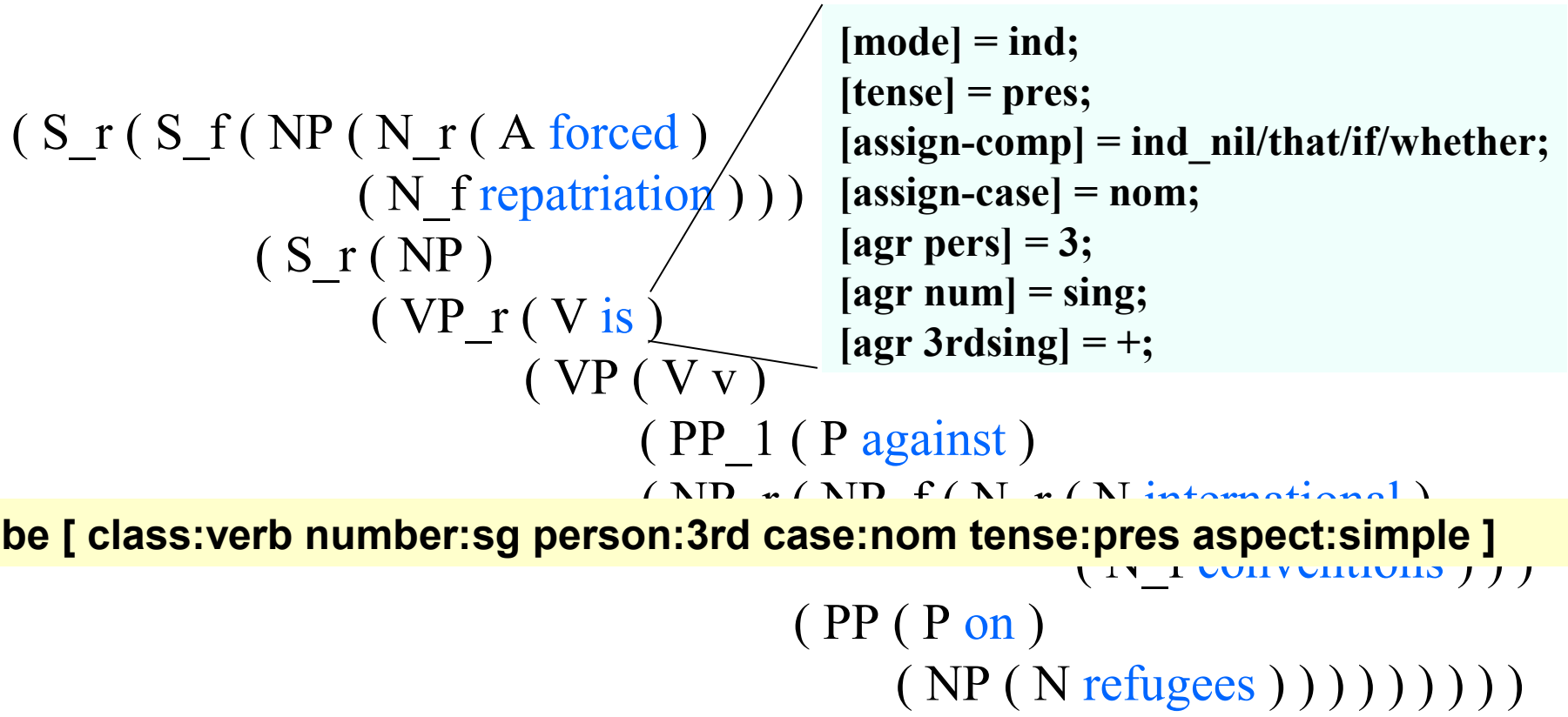
NLWM System I



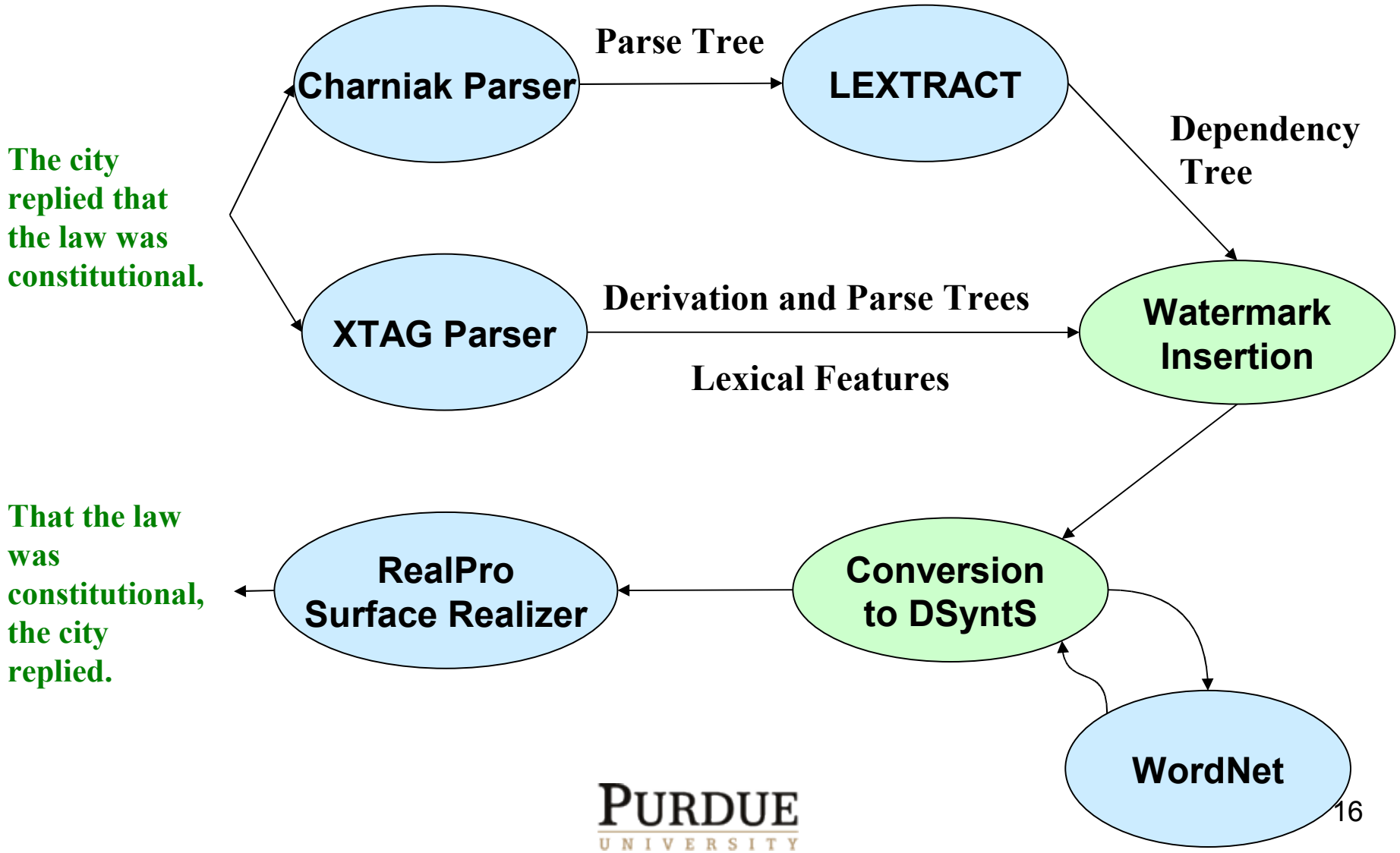
XTAG Parser

- **Lexicalized Tree Adjoining Grammar**
  - Features of nodes are unified through the levels of the tree

Forced repatriation is against international conventions on refugees.



NLWM System II



Evaluation of Natural Language Watermarking

- **Evaluating natural language watermarking**
  - User perception model is not as developed
  - Fluency is only preserved by sentence level changes
  - Style may be evaluated through authorship attribution techniques
  - Meaning and Grammaticality can be classified under *Adequacy*

Evaluating the Adequacy of Generated Text

- **How to measure the adequacy of generated text ?**
- **Criterion from MT: Number of n-grams that sentences generated by the system share with the cover sentences**
- **Differences exist from MT Evaluation**
  - Comparison among more than one generated texts
  - Human translations vs machine translation

Evaluation of Machine Translation

- **BLEU (BiLingual Evaluation Understudy) Metric**

(Papineni et al., 2002)

  - Based on n-gram matches

$$Score = \exp\left\{\sum_{n=1}^{N} w_n \log(p_n) - \max\left(\frac{L^*_{ref}}{L_{sys}} - 1, 0\right)\right\}$$

$$p_n = \frac{\sum_i \left(\begin{array}{l} \text{the number of n-grams in segment i, in the} \\ \text{translation being evaluated, with a matching} \\ \text{reference cooccurence in segment i} \end{array} \right)}{\sum_i \left(\begin{array}{l} \text{the number of n-grams in segment i, in the} \\ \text{translation being evaluated} \end{array} \right)}$$

$$w_n = N^{-1}$$

Evaluation of Machine Translation

- **NIST Metric** (Doddington, 2002)
  - Based on the informative value of n-gram matches

$$Score = \sum_{n=1}^{N} \left\{ \sum_{\substack{\text{all } w_1 \ldots w_n \\ \text{that co-occur}}} Info(w_1 \ldots w_n) \right\} \exp\left\{ \beta \log^2 \left[\min\left(\frac{L_{sys}}{L_{ref}}, 1 \right) \right] \right\}$$

$$Info(w_1 \ldots w_n) = \log_2 \left(\frac{\text{the number of occurences of } w_1 \ldots w_{n-1}}{\text{the number of occurences of } w_1 \ldots w_n} \right)$$

Results

- **Cumulative N-gram Scoring over 683 sentences**

NIST	1-gram	2-gram	3-gram	4-gram	5-gram
System I	7.3452	9.0238	9.2225	9.2505	9.2536
System II	6.2987	7.3787	7.4909	7.4962	7.4965

BLEU	1-gram	2-gram	3-gram	4-gram	5-gram
System I	0.8511	0.6694	0.5448	0.4532	0.3821
System II	0.7693	0.5096	0.3484	0.2439	0.1724

NIST 2005 Machine Translation Evaluation best score for BLEU 4 gram was 0.5137.

Discussion

- **Baseline coverage and quality tests**
- **NLWM System I**
  - Simpler
  - Less information feed into watermarking
  - Parser accuracy is low
  - Better BLEU score
- **NLWM System II**
  - Complex
  - More information feed into watermarking
  - Lower BLEU score
  - Needs a more sophisticated information management

Conclusion and Future Work

- **Many research and implementation challenges are involved**
  - Preserving meaning, fluency, grammaticality and style in text
- **Reasonable baseline text quality is achieved**
  - Our best system achieves a 0.4532 BLEU 4 score
- **Important step in mitigating security and privacy requirements of information exchange based on text**

Information Hiding in Natural Language

- **http://www.cerias.purdue.edu/homes/mercan**