Hiding the Message behind the Words: Advances in Natural Language Watermarking

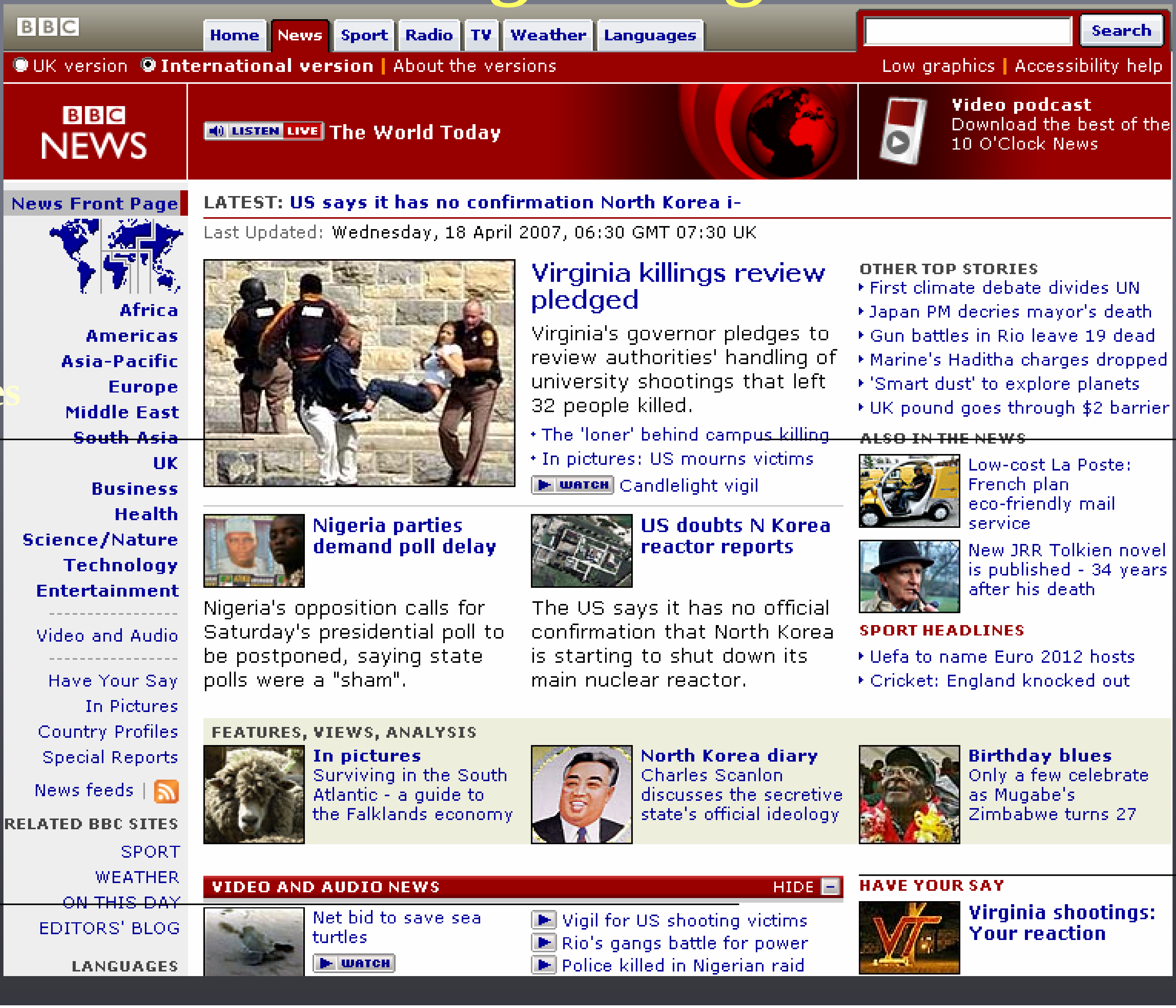
Mercan Topkara

Department of Computer Science

Purdue University

West Lafayette, 47907, IN, USA

Digital Age


BBC
Home
News
Sport
Radio
TV
Weather
Languages
Search
○ UK version ● International version | About the versions
Low graphics | Accessibility help
BBC NEWS
LISTEN LIVE
Video podcast
WATCH
HIDE

The World Today

Video podcast
Download the best of the
10 O'Clock News

News Front Page

Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
Business
Health
Science/Nature
Technology
Entertainment

Video and Audio

Have Your Say
In Pictures
Country Profiles
Special Reports
News feeds |

RELATED BBC SITES
SPORT
WEATHER
ON THIS DAY
EDITORS' BLOG
LANGUAGES

LATEST: US says it has no confirmation North Korea i-

Last Updated: Wednesday, 18 April 2007, 06:30 GMT 07:30 UK

Virginia killings review pledged

Virginia's governor pledges to review authorities' handling of university shootings that left 32 people killed.

- The 'loner' behind campus killing
- In pictures: US mourns victims
- WATCH Candlelight vigil

Nigeria parties demand poll delay

Nigeria's opposition calls for Saturday's presidential poll to be postponed, saying state polls were a "sham".

US doubts N Korea reactor reports

The US says it has no official confirmation that North Korea is starting to shut down its main nuclear reactor.

OTHER TOP STORIES
- First climate debate divides UN
- Japan PM decries mayor's death
- Gun battles in Rio leave 19 dead
- Marine's Haditha charges dropped
- 'Smart dust' to explore planets
- UK pound goes through $2 barrier

ALSO IN THE NEWS

Low-cost La Poste: French plan eco-friendly mail service

New JRR Tolkien novel is published - 34 years after his death

SPORT HEADLINES
- Uefa to name Euro 2012 hosts
- Cricket: England knocked out

FEATURES, VIEWS, ANALYSIS

In pictures
Surviving in the South Atlantic - a guide to the Falklands economy

North Korea diary
Charles Scanlon discusses the secretive state's official ideology

Birthday blues
Only a few celebrate as Mugabe's Zimbabwe turns 27

VIDEO AND AUDIO NEWS

Net bid to save sea turtles
WATCH

- Vigil for US shooting victims
- Rio's gangs battle for power
- Police killed in Nigerian raid

HAVE YOUR SAY

Virginia shootings: Your reaction

Images

Text

YOUR SAY

Audio

Video

Acknowledgements

- In different phases of the presented work, I have closely collaborated with (in alphabetical order):
    - Mikhail J. Atallah, Edward J. Delp, Dilek Hakkani-Tur, Victor Raskin, Giuseppe Riccardi, Cuneyt Taskiran, Umut Topkara
    - also briefly collaborated with Srinivas Bangalore and Owen Rambow

Problem

- **Controlling how the information we create is distributed or re-used**
  - How can you be sure that your articles/ papers/ blogs/ e-mails are not re-used without due credit?
  - How can you trust an email coming from the mail address of a person or an institute is really written by them?
- **Need for a rights protection system that travels with the content**
- **Approach: Information Hiding**

Glance at Information Hiding

- Information Hiding
  - Watermarking
    - Visible
    - Invisible
      - Fragile
        - Tamper Proofing
      - Robust
        - Fingerprinting
        - Meta-data binding
        - Copyright Protection
  - Steganography

THOMAS J. WATSON RESEARCH CENTER
IBM

PURDUE
UNIVERSITY

What is Natural Language Watermarking?

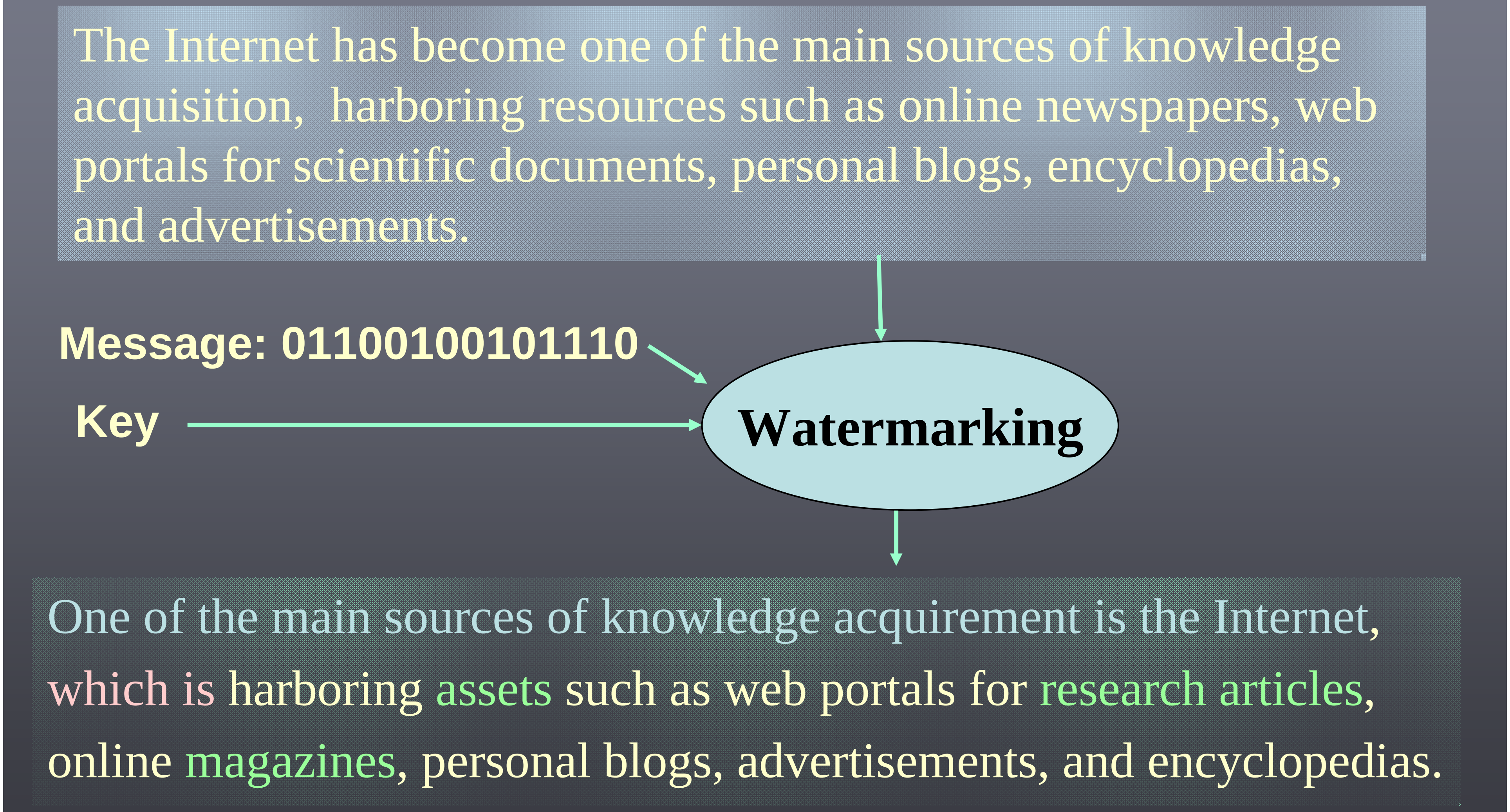
- **Enable copyright holders to enforce their intellectual property ownership on text**
- **Value of text:**
  - Meaning
  - Grammaticality
  - Style
- **Mark the text such that:**
  - The marking modifications do not reduce text’s value
  - Adversary will reduce text’s value to remove the mark

Why Natural Language Watermarking?
(Applications)

- Authenticating the source of a document
- Proving or denying ownership on a document
- Controlling distribution and reuse of intellectual property
- Digital libraries, on-line newspapers and stores etc.
- Content protection, text auditing, meta-data binding, tamper-proofing, traitor tracing, fingerprinting

An Example

How can we perform fully automatic robust watermarking?



The Internet has become one of the main sources of knowledge acquisition, harboring resources such as online newspapers, web portals for scientific documents, personal blogs, encyclopedias, and advertisements.

Message: 01100100101110

Key

Watermarking

One of the main sources of knowledge acquirement is the Internet, which is harboring assets such as web portals for research articles, online magazines, personal blogs, advertisements, and encyclopedias.

Traditional Challenges

- **Low bandwidth**
  - **Short documents**
  - **Not all transformations can be applied to a sentence**

	(I run by the river every morning.)
Grammar	(The river run me every morning.)
Meaning	(I manage by the river every morning.)
Style	(I don’t not run by the river every morning.)

Traditional Challenges

- Powerful Adversary
  - Can automatically edit individual sentences
  - Can permute sentence order
  - Can delete or insert sentences
  - Has access to the same data and software resources

NLP for Watermarking

- Natural Language Processing (NLP) aims to design algorithms that will analyze, understand, and generate natural language automatically
- Electronic Data Resources and Tools
  - Corpora
  - Dictionaries e.g., WordNet, Verbnet
  - Parsers, Generators, Machine Translation and Question Answering Systems

Previous Approaches

- **Generating the cover text (Steganography)**
  - Passive Adversary
  - Cover text has no “value”
    - Spammimic (M. Chapman and G. Davida, 2002)
- **Modifying a given cover text**
  - Active Adversary
  - Proposed for steganography as well as watermarking

Previous Work in Linguistic Steganography

- **Mimicry Text: Using PCFGs to Generate Cover Text**

(Wayner, 1992)

Rule	Code	Prob.
S → AB	0	0.5
S →CB	**1**	**0.5**
A → She	00	0.25
A → He	01	0.25
A → Susan	10	0.25
A → Alex	11	0.25
B → likes D	0	0.5
B → detests D	10	0.25
B → wants D	**110**	**0.125**
B → hates D	111	0.125
C → Everybody B	**0**	**0.5**
C → The lady B	10	0.25
C → A nice kid B	11	0.25
D → milk.	00	0.25
D → apples.	01	0.25
D → pie.	**10**	**0.25**

Position	Prefix	Output
°1011010	1	CB
1°011010	0	Everybody B
10°11010	110	Everybody wants D
10110°10	10	Everybody wants pie.

Previous Work in NL Watermarking

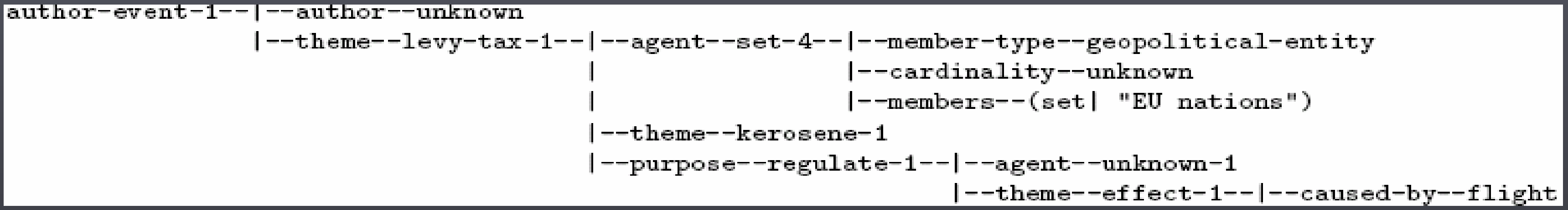
- **Encode the bit string in the tree structure**
- **Change the tree with Syntactic Transformations**

(Atallah, Raskin et. al, 2001)

```
(S (NP Ned)
   (VP loves (NP Jody))
   (. .))
```

- **Change the tree with Semantic Transformations**

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
```

M. Atallah, V. Raskin, C. F. Hempelmann, M. Karahan, R. Sion, U. Topkara, K.E. Triezenberg, "Natural Language Watermarking and Tamperproofing", IHW 2002

Our Approaches

- **How can you provide resilience to removal attacks?**
  - By hiding the information carriers
    - Sentence Level Watermarking
  - By making it very hard to undo the embedding changes
    - Embedding through the use of ambiguity

Sentence Level Watermarking: Enigmark

- **Linguistic transformations are defined at sentence level as opposed to individual words**
- **Provides a large feature space**
  - words, phrases, punctuation, parse structure, etc.
- **Sentence level watermarking using multiple orthogonal features (Enigmark)**
- **Selection is orthogonal to embedding**

This frank discussion will close this chapter.

↓

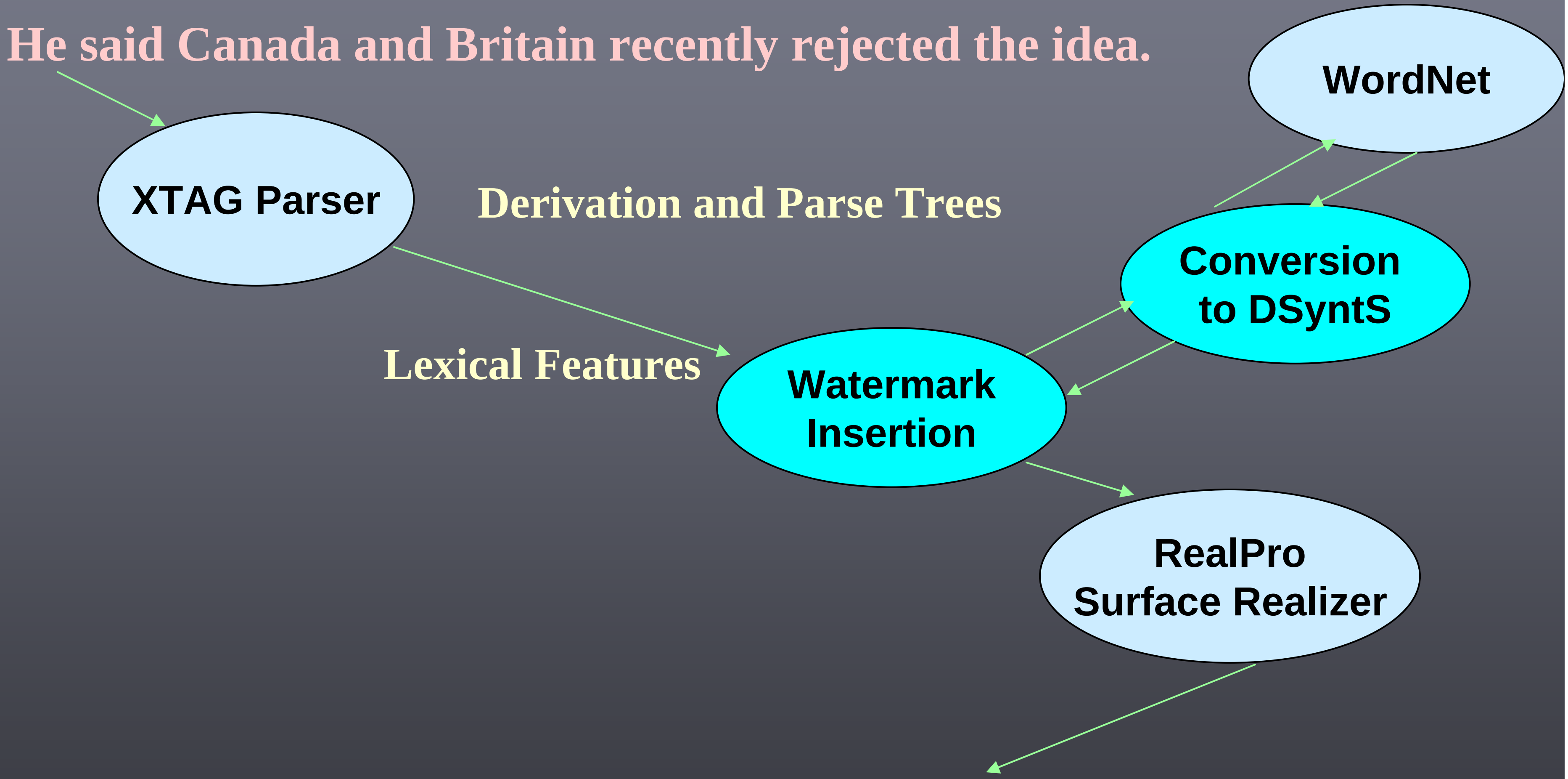
This chapter will be closed by this frank discussion.

This chapter, by this frank discussion, will be closed.

Sentence Level Watermarking

- **Robust features**
  - Hard to undo
  - Used to select the watermark carrying sentences
  - E.g. verb classes, ambiguous words, etc.
- **Yielding features**
  - Easy to change
  - Used to embed the watermark bits
  - E.g. structure, punctuation, phrases, word order, etc.
- **M. Topkara, U. Topkara, M. J. Atallah, "Words Are Not Enough: Sentence Level Natural Language Watermarking", ACM MCPS'06.**

Enigmark: System Snapshot



He said Canada and Britain recently rejected the idea.

XTAG Parser

Derivation and Parse Trees

Lexical Features

WordNet

Conversion to DSyntS

Watermark Insertion

RealPro Surface Realizer

He said the idea was recently rejected by Canada and Britain.

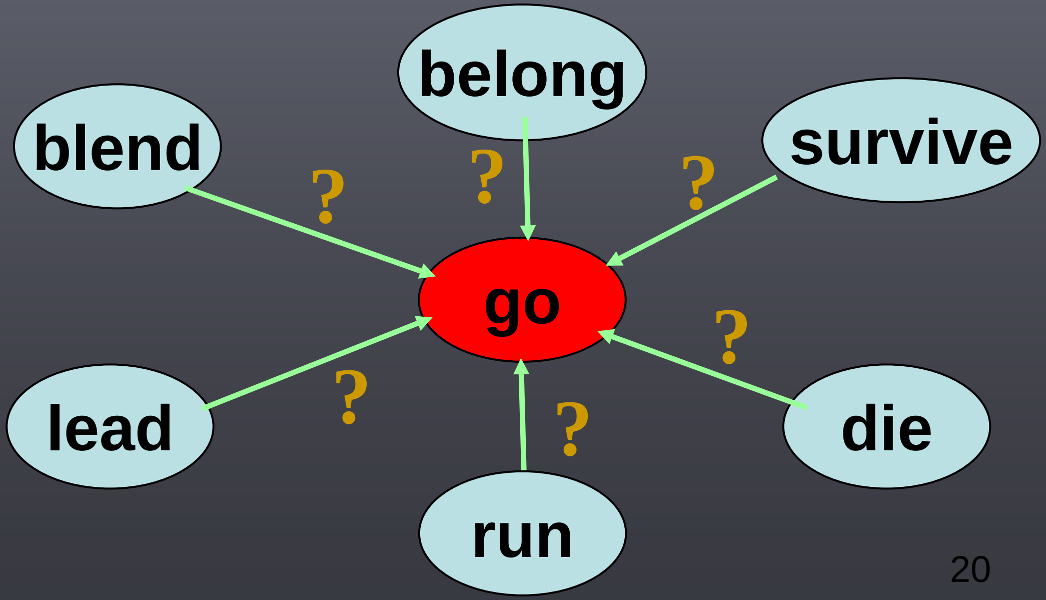
Our Approaches

- **How can you provide resilience to removal attacks?**
  - By hiding the information carriers
    - Sentence Level Watermarking
  - By making it very hard to undo the embedding changes
    - Embedding through the use of ambiguity

Hiding Virtues of Ambiguity

- **Goal:** Lowering adversary's power
- **Approach:** Increasing the adversary's uncertainty
  - The adversary is a machine
- **Means:** Knowledge asymmetry between the embedding process and the adversary

Go Boilers!!!!!



Computationally Asymmetric Transformations

- **Can be carried out inexpensively**
- **Yet reversal requires disproportionately larger computational resources or human intervention**
- **Robust synonym substitution (Equmark)**

This robot is very smart. (Original)

↓

This robot is very bright. (Modified)

Is the robot polished or smart?

- **U. Topkara, M. Topkara, M. J. Atallah, "The Hiding Virtues of Ambiguity: Quantifiably Resilient Watermarking of Natural Language Text through Synonym Substitutions" , ACM MMSEC’06.**

Equmark: Technical Details

- **Build a graph, G, of (word, sense) pairs**
  - WordNet
- **Assign weights to the edges**
  - Using a “word similarity measure”
- **Select a sub-graph, G^W, of G using a secret key, k**
- **Color G^W using k**
  - 3 different colors are used to assign “0”, “1”, “no-encoding” to the words
  - Homographs in the same synonym set get opposite colors

Equmark: Quantifying Distortion

- Watermark embedding distortion

$$\sum_{s^N \in S^N} \sum_{k^N \in K^N} \sum_{m \in M} \frac{1}{|M|} p(s^N, k^N) d_1^N(s^N, f_N(s^N, m, k^N)) \leq D_1$$

- Maximum distortion an adversary can introduce

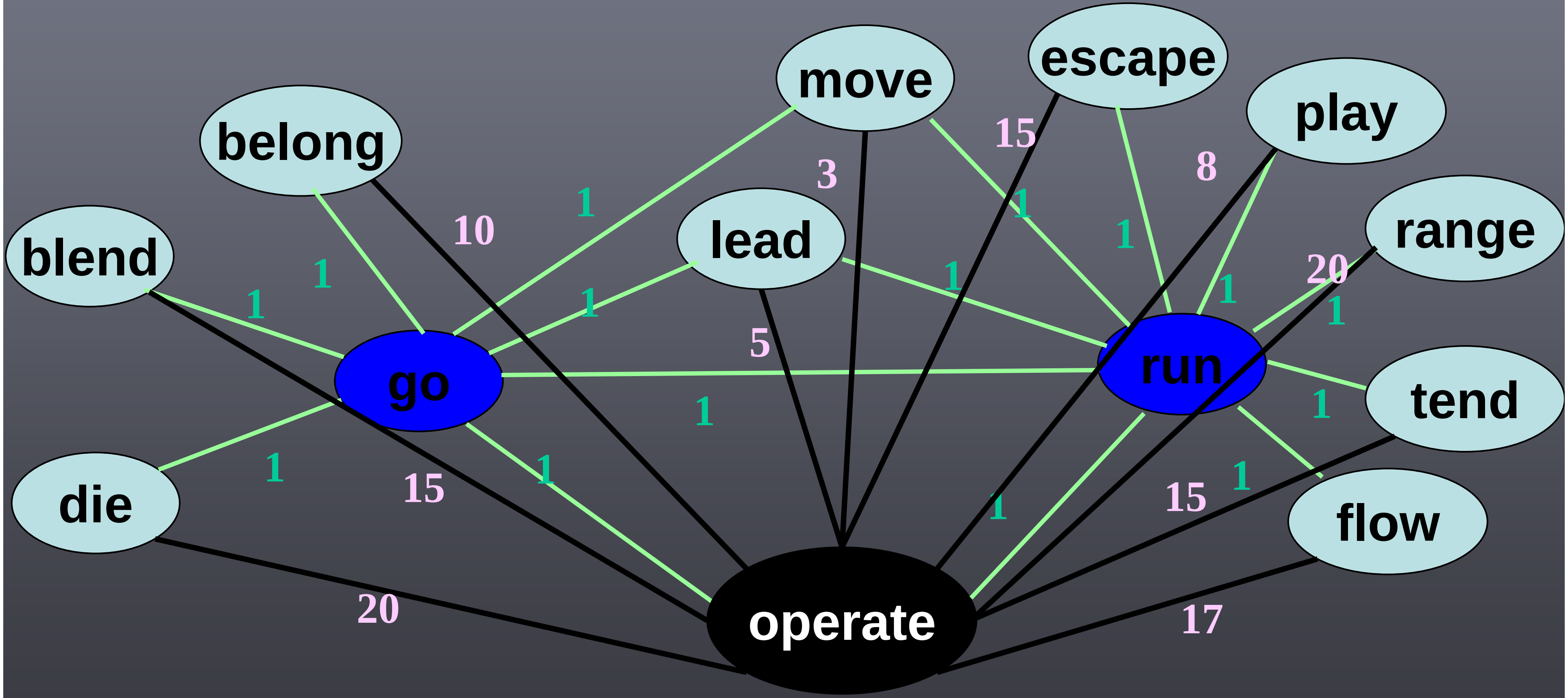
$$\sum_{x^N \in X^N} \sum_{y^N \in Y^N} d_2^N(x^N, y^N) A^N(y^N \mid x^N) p(x^N) \leq D_2$$

- When there are more than one alternatives pick the one that stays below the embedding distortion while maximizing expected distortion of the adversary

Equmark: Quantifying Distortion

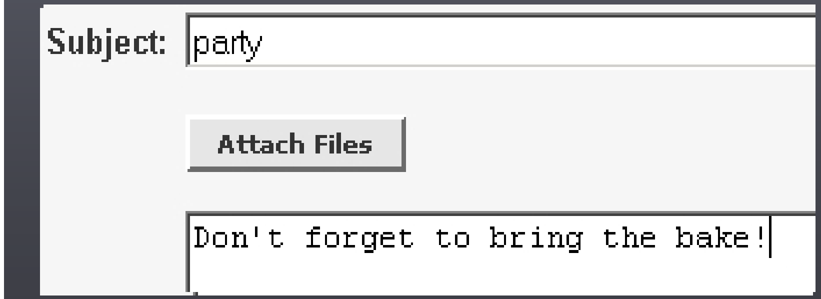
(Mock Example)

- Maximize the expected distortion of the adversary
- Similarity of synonyms is 1

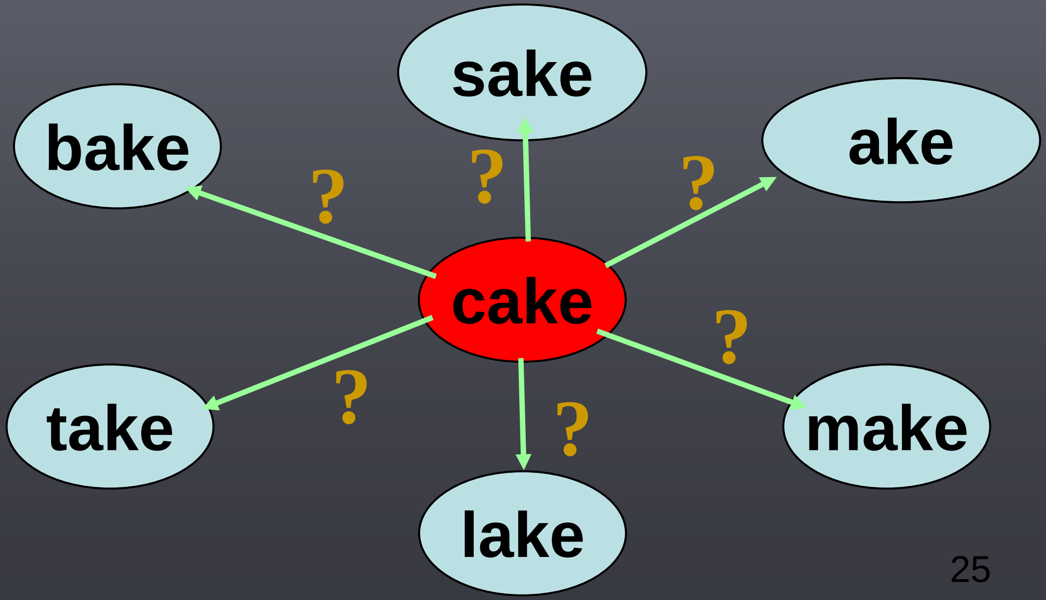


Marking Cursory Text

- **Goal: Marking a different genre: cursory text**
  - E-mails, blog entries, forums etc.
- **Approach: Increasing the adversary’s uncertainty**
  - **Adversary is a machine and reader is a human**
  - **Humans are good at spelling correction**
- **Means: Typographical errors [SPIE 2007]**


Subject: party

Don't forget to bring the bake!


cake

sake ?

bake ?

ake ?

take ?

lake ?

make ?

Marking Cursory Text: Markerr

- **The substantial portion of the daily exchanged text**
  - Emails, text messages, forum posts, blogs, wikis.
- **Typographical errors (typos) are common**
- **Typos can occur in any part of the text**
- **Humans adapt to errors in these type of text, and are good in spelling correction**
- **Idiosyncrasies of cursory text create room for information hiding**
  - **“teh”, “lol”, “l33t”, “:)”, “gonna”, “DCT”**

Markerr: Case for Ambiguity

- **Challenge: Powerful spell checkers**
- **Approach: Use ambiguous or stealthy typos**
  - **Typos that are close to many words**

    “world” → “worod” (was it wood, word or world?)
  - **Word to word conversions**

    Don’t forget to bring the cake.

    “cake” → “sake” (looks correct)
  - **Acronyms have two-way ambiguity**

    “gbh” → { “great big hug”, “grievous body harm”}

    “good to see you” → {“GTCY”, “GTSY”, “G2CY”, “G2SY”}

Markerr: Case for Ambiguity

- Different models against the adversary
  - Maximizing the adversary’s uncertainty about the original word

$$h(\overline{w}) = -\sum_{a\in N(\overline{w})} \Pr(a\,|\overline{w})\log(\Pr(a\,|\,\overline{w}))$$

  - Maximizing the probability that the inserted bit will stay the same even if the adversary randomly updates the message carrying word

$$\Pr(m_i\,|\,\overline{w}) = \sum_{a\in N'(\overline{w})} \Pr(m_i\,|a)$$

- **M. Topkara, U. Topkara, M. J. Atallah, "Information Hiding through Errors: A Confusing Approach", SPIE 2007.**

Impact and Conclusions

- Presented three different schemes
  - Sentence level watermarking (Enigmark)
  - Robust synonym substitution (Equmark)
  - Ambiguous errors (MarkErr)
- Cover several types of text
  - Short, long, edited, cursory…
- Provide light mark reading process
  - Inexpensive language analysis at the detection time

Impact and Conclusion

- **Challenging problem**
  - **Natural language text as a cover media**
- **Wide range of application areas**
  - Content protection, meta-data binding, tamper proofing, fingerprinting…
- **Increased interest in the research area**

Future Work

- Applications
- Evaluation techniques
- Large scale user study
- Increasing capacity and resiliency
- Different languages, different genres