Natural Language Processing: Some *wh*-questions

Laurette Pretorius
School of Computing

October 22, 2007

UNISA university of south africa

Introduction

- Good news flash should address at least the wh-questions *What? Who? Where? When?* and *Why?*

- The *wh*-questions on NLP: What? Why? and How?

What?

- NLP: design and implementation of computer technology that is able to communicate with humans in *natural language*
- Ideal: Language input-output components of artificially intelligent systems that are capable of using language as fluently and flexibly as humans do (analyse, understand and generate)
- Linguistics 100 jaar
- Computer science and NLP 50 jaar
- Related disciplines:

What? (cont.)

- Cognitive science: explains the phenomenon of human language
- Generative linguistics: formalises human language
- AI: simulates intelligent behaviour (including human language)
- Computational linguistics: implement language theories, natural vs. formal languages, electronic corpus analysis
- Electrical, electronic and computer engineering: signal processing, speech recognition, speech synthesis
- Computer science: “tools” (theory and practice) for realising NLP

Why?

- Natural language (spoken and written): preferred mode of communication, also electronic communication
- Information explosion: retrieval, extraction, QA systems, text summarisation, terminology extraction, MA translation, MT of text and speech, web developments, …
- Examples (use of knowledge of language central): word processing, computer games, CA education, dialog systems and HCI, autonomous communicating agents, …

Why? (cont.)

- Cultural diversity
- Endangered languages: describe, archive, empower, preserve …
- Empowering marginalised language communities
- Rich domain of application of everything that computer science has to offer
- 1950s, 1980s, 1990s, sedert 2000 …

How?

Challenge: Robust computational resolution of ambiguity in natural language w.r.t.

- sounds (phonetics and phonology)
- words (morphology)
- sentences (syntax)
- meaning (semantics)
- discourse context (pragmatics and discourse), ...

Approaches:

- Rule- or knowledge-based
- Statistical and probabilistic
- Machine learning techniques

How? (cont.)

Formal models:

- State machines, including FSAs, FSTs, weighted automata, HMMs
- Formal rule systems, including regular grammars, CFGs, constraint-based or feature-augmented grammars
- Logic and the large industry of applied logics
- Probability theory and machine learning

First two: phonology, morphology and syntax
Second two: semantics, pragmatics and discourse

Other essentials:

- Electronically available language resources
- Scientific evaluation and standards
- Applications

UNISA university of south africa

More specifically …

Unisa, collaborating with a variety of linguists:

- Zulu, Xhosa, (Swati, Ndebele) and Tswana: Morphological analysis
- Afrikaans: Shallow parsing
- Venda, Malagassy, Khoekhoegowab: Students
- Northern Sotho: Talking head
- Machine-readable lexicon development
- Computational tools for exploring the African languages spoken language corpora
- Ontologies, Wordnets and the Semantic Web in the context of the South African languages

From components to real applications and solutions …

UNISA university of south africa