Computational Linguistics

Introduction

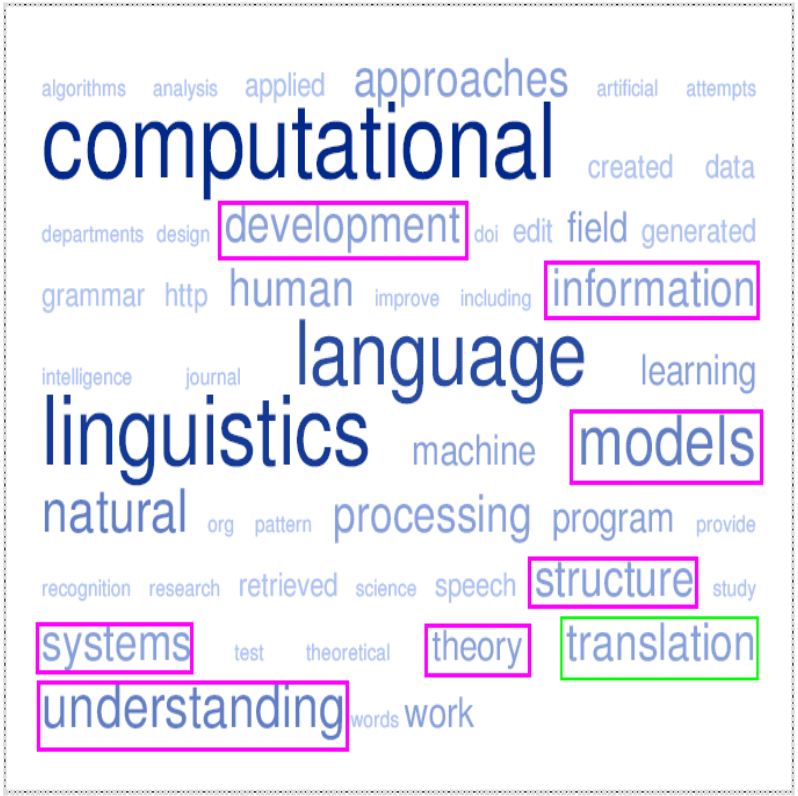
Bali Ranaivo-Malançon & Suhaila Saee

Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak

August 2014

Computational Linguistics (CL)

- Classified as APPLIED LINGUISTICS
- Makes use of computer to develop formal systems that model human languages for language understanding
- Often considered as synonym to NATURAL LANGUAGE PROCESSING



(Word cloud generated by *TagCrowd* (http://tagcrowd.com/) from "Computational Linguistics" article in *Wikipedia*)

Natural Language Processing (NLP)

- Classified as a sub-field of ARTIFICIAL INTELLIGENCE
- Concerns with the design, implementation, and evaluation of algorithms to process human languages for human-machine interaction

(Word cloud generated by *TagCrowd* (http://tagcrowd.com/) from "Natural Language Processing" article in *Wikipedia*)

Language Technology (LT)

- aka **Human Language Technology**
- Very close to Language Engineering
- Concerns with practical applications that deal with human languages and have a real impact on human lives and business

(Word cloud generated by *TagCrowd* (http://tagcrowd.com/) from "What is Language Technology?" by Hans Uszkoreit (http://www.dfki.de/lt/lt-general.php))

Ideal Picture!

A computer that can understand and generate natural language

Linguistics

- The scientific study of human languages (or natural languages)
- It includes the study of language from prescriptive, comparative, structural, and generative points of view

(Word cloud generated by *TagCrowd* (http://tagcrowd.com/) from "Linguistics" article in *Wikipedia*)

Linguistic Components & Units

Linguistic Component	Definition	Linguistic Units
Phonetics	The study of speech sounds.	Phonetic features
Phonology	The study of sound systems.	Phonemes; Syllables
Morphology	The study of word formation and structure.	Morphemes
Syntax	The study of sentence structure.	Phrases; Clauses; Sentences
Semantics	The study of meaning of (part of) words, phrases, sentences, and texts	Semantic features
Pragmatics	The study of the aspects of meaning and language use.	Utterances
Discourse	The study of text structure and function of text and conversation	Discourse; Turns

Characteristics of Natural Languages

- The most important means of human communication
- Realised
  - acoustically (sound waves)
  - visually-spatially (sign language)
  - in written form (writing systems)
- **Complex**, **ambiguous**, and **noisy**
- Different from
  - Formal languages (e.g. programming languages, mark-up languages, etc.)
  - Constructed languages (e.g. Esperanto, Glosa, Ido, Interlingua, etc.)

Many Natural Languages!

Around 6,900 known living languages in the World

Area	Living languages		Number of speakers			
	Count	*Percent*	*Count*	*Percent*	*Mean*	*Median*
Africa	2,110	30.5	726,453,403	12.2	344,291	25,200
Americas	993	14.4	50,496,321	0.8	50,852	2,300
Asia	2,322	33.6	3,622,771,264	60.8	1,560,194	11,100
Europe	234	3.4	1,553,360,941	26.1	6,638,295	201,500
Pacific	1,250	18.1	6,429,788	0.1	5,144	980
Totals	6,909	100.0	5,959,511,717	100.0	862,572	7,560

(Source: *Ethnologue*, http://www.ethnologue.com/)

Writing Systems

- aka **scripts**
- Represent linguistic units at different structural levels
- Each script has
  - its own set of icons to represent characters (or letters) and numerals
  - its own set of rules to combine the characters
- In addition, there are icons for special symbols found on keyboards

Many Writing Systems!

- **Alphabetic** - Consonants and vowels are symbolised; e.g. French, English
- **Logographic** - Every symbol represents a word or morpheme; e.g. Chinese
- **Syllabic** - Each syllable is represented by its own symbol, and words are written syllable by syllable; e.g. Japanese
- **Consonantal alphabet** - Each symbol represents a consonant and vowels may be represented by diacritical marks; e.g. Semitic languages like Arabic, Hebrew
- **Emoticons** - Graphic symbols representing facial expressions and used to show someone's feeling

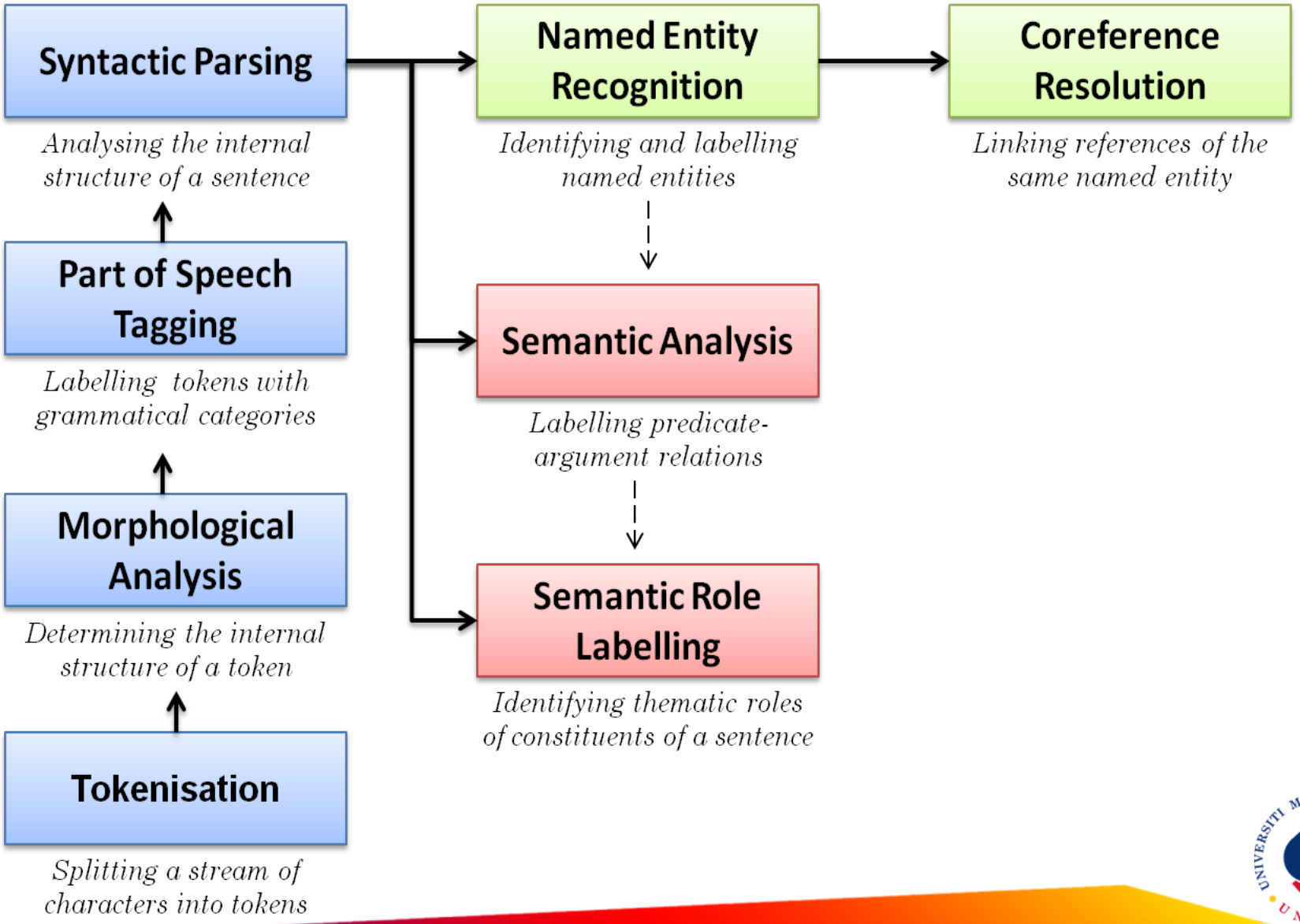
Language Resources

- Each CL/NLP application requires specific language resources
- Language resources vary from simple wordlists to complex resources like ontology
- CL/NLP provides different methods for creating, recording, processing, and reusing language resources

Kinds of Language Resources

- Lexicon or machine readable dictionary
- Terminological resources: thesaurus, term bank
- Ontology or knowledge base
- Translation memory
- Written or spoken corpus

Generic CL/NLP Pipeline



Symbolic CL/NLP Approaches

- `Symbolic`: a natural language is considered as a sequence of symbols
  - A written text is a sequence of sentences, a sentence is a sequence of words, and a word is a sequence of letters
- Considered as **classical approaches** by the authors of the second edition of the *Handbook of Natural Language Processing*
- Characterised by the use of a large set of hand-crafted rules to encode linguistic symbols
- Thus the name, **rule based approaches**

Empirical & Statistical Learning CL/NLP Approaches

- `Empirical`: knowledge is derived from observation (rather than theory)
- `Statistical`: the observed data is transformed into a probabilistic model
- `Learning`: general rules are inferred by observing examples
- Characterised by the use of a large collection of documents (written or spoken corpus) to derive language models from statistical methods
- Thus the name, corpus based approaches

Connectionist CL/NLP Approaches

- **Connectionism** provides a framework for the study of cognition using artificial neural network (ANN) models
  - ANN = neural network = neural net
  - ANN models consist of a set of identical processing units called artificial neurons
  - Artificial neurons are interconnected via weighted connections
- Attempt to model human intellectual abilities using ANNs as non-symbolic systems
- Many learning algorithms exist to implement learning in ANNs
  - A learning algorithm determines appropriate changes in the weight values to perform a set of input/output mappings

Some CL/NLP Applications

- Machine translation
  - To translate automatically a source language into a target language
- Information retrieval
  - To find relevant documents within a large amount of unstructured data collection through keyword-queries formulated by a user
- Question answering
  - To find relevant short answers within a large amount of unstructured data collection through natural language questions formulated by a user

Some CL/NLP Applications (cont'd)

- Text summarisation
  - To generate an abstractive or extractive summary of a single document or a set of documents
- Spelling checker / Grammatical checker / Style checker
  - To correct the spelling/grammar/style errors within a written document
- Sentiment analysis / Emotion recognition / Opinion mining
  - To determine opinion/emotion expressed in unstructured natural language texts

Some CL/NLP Tools

- NLTK (Natural Language ToolKit)
  - An open-source project that offers a large collection of Python modules to process natural languages
- GATE (General Architecture for Text Engineering)
  - An open-source platform written in Java for developing and deploying software components that process human languages
- OpenNLP
  - A collection of Java-based NLP tools
- LingPipe
  - A toolkit for processing text using computational linguistics
- UIMA (Unstructured Information Management Architecture)
  - An open-source framework by IBM for building content analytic applications