Introduction to the Natural Language Processing (NLP)

Thierry Hamon
thierry.hamon@univ-paris13.fr

LIM&BIO – EA 3969
Université Paris 13 - UFR Léonard de Vinci
74, rue Marcel Cachin, F-93017 Bobigny cedex
http://www-limbio.smbh.univ-paris13.fr/membres/hamon/

February 2013
Plan

- History and context
- Example
- Introduction to NLP approaches
- Formal language vs. Natural language
- Evaluation
The very beginning

- Context: Back in the fifties (cold war)
- Main application: Machine translation
  use of computers to translate texts or messages from one (source) language to a other language (target language)
- Budget: about $20 millions in 10 years
The mythological tests/jokes

- Translation of the Biblical sentence

  *The spirit is willing, but the flesh is weak*
  
  or
  
  *Out of sight, out of mind*

- Translation in Russian, and then in English
The mythological tests/jokes

- Translation of the Biblical sentence

  \textit{The spirit is willing, but the flesh is weak}
  
or
  \textit{Out of sight, out of mind}

- Translation in Russian, and then in English

  \textit{The vodka is strong, but the meat is rotten}

  \textit{Invisible idiot}
The mythological tests/jokes

- Translation of the Biblical sentence
  
  *The spirit is willing, but the flesh is weak*
  
  or
  
  *Out of sight, out of mind*

- Translation in Russian, and then in English
  
  *The vodka is strong, but the meat is rotten*

  *Invisible idiot*

Literal translation (word for word translation) is inappropriate
(problem with idioms)
More information is needed
The linguistic underside

- Machine readable dictionaries
- Syntactic information (order and function of the words)
- Problems:
  - Ambiguities, polysemy, ...
  - Complex syntactic structures,
  - Semantics (relations, categories, ...)
  - Anaphora, ...
The (in)famous "ALPAC report"

- In 1966, by the US National Academy of the Sciences
  Y. Bar-Hillel
  - Complete machine translation: slow, time consuming, with a low quality
  - could be more expensive than human translators
  - *Machine Translation is hopeless* (!)
  - Recommendations:
    - Evaluation of the translations (quality and cost)
    - Machine-aided translation
    - More efforts on the computational linguistic research
    - For machine translation or not

- Conclusion; lower budget for machine translation but the beginning of the Natural Language Processing (NLP)
Contributions

Interdisciplinary research field:

- Mathematics:
  - Logic
  - Formal language theory
  - Statistics

- Computer science
  - Algorithms
  - Software engineering
  - Machine learning

- Linguistics
  - Phonology
  - Generative grammars
  - Syntax
  - Philosophy of language
Research fields

Two main fields

- 1960 Computational linguistics
  Focus on mathematics and linguistics

- 1965 Natural Language Processing
  Focus on algorithms for software development

1970 Natural Language Understanding (AI)
Cognitive approaches

- T Winograd, M Minski, J Allen, ...
50 years later

- Phonetics, phonology, prosody
- Morphology
- Syntax
- Semantics
- Pragmatics
## Introduction to NLP approaches

### Phonetics
- Pronunciation
- Syllabation
- Prosody
- lexique.org, ...

### Morphology
- Inflected form
- Derivation
- Composition
- MorTAL, Celex, ...

### Syntax
- Syntactic lexicon
- Syntactic analysis
- Chunking
- LTAG, FTAG, LFG, ...

### Semantics
- Semantic network
- Semantic lexicon
- Terminology
- WordNet, DEC, ...

### Pragmatics
- Desambiguation rules
- Extraction of semantic units
- Relation acquisition
- Decomposition en primitives
- Definition analysis

### Resources
- Text structure
- Anaphora
- Communication

### Tasks
- Speech Recognition
- Speech synthesis (text speech)
- Part–of–speech tagging
- Syntactic analysis
- Chunking
- Morphological segmentation
- Morphological analysis

### Applications
- Speech recognition
- Spell checking
- Man machine dialogue
- Corpus Linguistics
- Resource building
- Text Generation

- Weather forecast, report, ...
- Stylistics
- Statistical NLP
- Automatic summarization
- MT (Machine Translation)
- CAT (Computer–assisted Translation)
- IR (Information Retrieval)
- IE/TM (Information Extraction/Text Mining)
- QA (Question Answering)

---

### History and Context

50 years later

---

**T Hamon**
Around the world

- ACL: Association for computational linguistics
- Journals: Computational LinguisticsL, JNLE, ...
- Conferences: ACL, COLING, EACL, NNACL, LREC, ...
- Web site: http://www.aclweb.org
- Mailing list: linguist, corpora
- Universities and research centers (JRC in Europ)
- Compagnies (Xerox, IBM, Microsoft, Lingeofter, etc.)
How to deal with the processing of natural language data?

- Natural language: system composed of signs, used to produce a utterance
- Words are basic signs of a language
- A word is made of two sides
  - Phonologic form (the signifier – train)
  - Meaning (the signified - the mental picture of the train)

(Ferdinand de Saussure, Cours de linguistique générale, 1916)
How to deal with the processing of natural language data?

- Several types of linguistic information help to go from one side to the other
- Those types of linguistic information are more or less autonomous
  Each interacts with others
How to deal with the processing of natural language data?

**Example**

*Query to a kiosk to get train schedule (by the mean of human speech)*

Location: Västerås Station

Question: *What time is the first train to Stockholm, tomorrow morning?*
How to deal with the processing of natural language data

First step

*Speech processing and recognition*

Computing of the speech signal to the words of the question

(Phonetics and phonology)
Phonetics and Phonology

- **Phonetics**: study of the sound of human speech (*phones*)
  - From the physical point of view
  - More related to Speech processing

- **Phonology**: Study of the groups of sound to make words or utterances in a natural language
  - From the linguistic point of view (*phonemes*)
  - Organisation of the sounds, syllabs, rhymes, etc.
  - Related to the meaning

The both also include the study of sign languages
How to deal with the processing of natural language data

Second step

Morphological analysis

- Description of the words regarding their form (*morpheme*)
- Recognition of the
  - Canonical form (dictionary entry)
  - Inflectional parameters (gender, number, ...)
  - Part of speech (grammatical category)

What is the first train to Stockholm, tomorrow morning?
How to deal with the processing of natural language data

Third step

Parsing (syntactic analysis)

- Combination of the words to make sentences
- Two points of view:
  - Recognition of
    - The constituents of the sentence (noun phrases, verbal phrases, adjectival phrases, ...)
    - The dependency between the words (modifier of a noun, subject of a verb, ...
How to deal with the processing of natural language data

Third step

(output of the Stanford parser)

det(time-2, What-1)
attr(is-3, time-2)
det(train-6, the-4)
amod(train-6, first-5)
nsubj(is-3, train-6)
prep_to(train-6, Stockholm-8)
nn(morning-11, tomorrow-10)
appos(Stockholm-8, morning-11)
How to deal with the processing of natural language data

Third step

(output of the Stanford parser)

(ROOT
  (SBARQ
    (WHNP (WDT What) (NN time))
    (SQ (VBZ is)
      (NP
        (NP (DT the) (JJ first) (NN train))
        (PP (TO to)
          (NP
            (NP (NNP Stockholm))
            (, ,)
            (NP (NN tomorrow) (NN morning))))))
  (? ?))))
How to deal with the processing of natural language data

Fourth step

**Semantic analysis**

- Identification of the
  - meaning of the words or phrases
  - semantic relations between them
- Without taking into account the context
- Logic can be used to represent semantics of a sentence
How to deal with the processing of natural language data

Fourth step

train → object, mode of transportation
first → first answer
first train?
Stockholm → Location/City/railway station/direction/destination (Stockholm C)
What time → Hour?
Tomorrow → (next day) Today + 1 day (27th of February, 2013)
morning → (daytime, day period) 8H00-12h00? 7H00-12H00? before noon?
...

How to deal with the processing of natural language data

fifth step

*Pragmatics*

- Semantic interpretation of the sentence according to the context

Contextual information:
- departure? (Västerås - Västerås C)
- date (today)? (26th of February, 2013)
- the results are sort by time (of departure)
- need of the schedule

but also, reference resolution (anaphora)
How to deal with the processing of natural language data

fifth step

Translation in SQL query: ad-hoc methods or compilation methods

```sql
SELECT MIN (startHour) FROM train
WHERE
departureDay = '02/27/2012'
AND departureLocation = 'Västerås'
AND arrivalLocation = 'Stockholm'
AND departureHour < 12:00 AND departureHour > 7:00;
```

(The answer is 7:10)
How to deal with the processing of natural language data

Comments

In the real, the kiosk could need more information (need of a human/machine dialogue)

What I didn’t say/ask (yet?):

- Direct train
- Track (at Västerås C and/or Stockholm C)
- Travel time
- Class
- Buy a ticket
- Return ticket
- Price
- Rebookable or not, Refundable
- For adult or child
- Number of tickets
How to deal with the processing of natural language data 
here and back again

**Answer generation**

- Translation of the query result into a textual form
  
  *The first train to Stockholm is at 7:10, tomorrow*

- In case of spoken answer, speech synthesis of the text
How to deal with the processing of natural language data?

Two directions:

- Analysis of language data (textual data or human speech) towards (more or less) the understanding of the message
- Generation of language data (textual data or speech synthesis) towards a linguistic realisation

Usually, NLP deals with the sentences
How to deal with the processing of natural language data?

Two paradigms for processing texts:

- **Symbolic paradigm**: extraction of linguistic information with symbolic information or linguistic resources
  Use of dictionary, grammars, rules

- **Stochastic paradigm**: use of stochastic approaches to extract linguistic information from textual corpora
  Use of machine learning (classification, decision trees, ...)

The both can be mixed
Presentations of NLP approaches

- Focus on NLP for acquisition and text understanding
- More or less with symbolic approaches
- Use of electronic texts: collection of textual documents
Which documents?

Texts or collection of texts: textual corpora

Great variations:

- Electronic formats (raw text, HTML, XML, PDF, Word, etc.)
- Character encoding (ASCII, ISO-LATIN-1, windows-1252, UTF-8, etc.)
- Type of documents (web pages, blogs, scientific articles, journal articles, books, tables, support group messages, emails, sms, etc.)
- Size: from few Kilo-bytes to several Giga-bytes

→ Work on raw text
Comparative Analyses of Hairpin Substrate Recognition by Escherichia coli and Bacillus subtilis Ribonuclease P Ribozymes.

Ando T, Tanaka T, Kikuchi Y.

Division of Bioscience and Biotechnology, Department of Ecological Engineering, Toyohashi University of Technology.

Previously, we reported that the substrate shape recognition of the Escherichia coli ribonuclease (RNase) P ribozyme depends on the concentration of magnesium ion in vitro. We additionally examined the Bacillus subtilis RNase P ribozyme and found that the B. subtilis enzyme also required high magnesium ion, above 10 mM, for cleavage of a hairpin substrate. The results of kinetic studies showed that the metal ion concentration affected both the catalysis and the affinity of the ribozymes toward a hairpin RNA substrate.
<h1 id="sidrubrik">About me - Sergei Silvestrov</h1>

<p>I am Professor in Mathematics and Applied Mathematics</p>

<p>I am also the Subject Chair for the subject of Mathematics and Applied Mathematics in Mälardalen University</p>

<p>&lt;a href="/polopoly_fs/1.49763!/CV%20for%20Sergei%20Silvestrov%201%20sida.pdf"&gt;CV (short, 1 page, pdf)&lt;/a&gt;</p>

→ Need to extract the content
About me - Sergei Silvestrov

I am Professor in Mathematics and Applied Mathematics

I am also the Subject Chair for the subject of Mathematics and Applied Mathematics

in Mälardalen University

(Ämnesföreträdare In Swedish)

CV (short, 1 page, pdf)
Pre-processing

- Cleaning of the texts (HTML markup)
- Homogenisation of the encoding charset
- Extra-linguistic normalisation
  - duplicated blank characters
  - hyphenation
  - font marks
  - typographic ligatures: difference, specific
  - long dash: – (--)
To be continued...
Formal language vs. Natural language

Formal language $L$:

- (possibly infinite) set of words $\Sigma^*$
- over a finite alphabet of symbols $\Sigma$
- word: finite sequence of symbols of the alphabet
- (syntactic) rules used to decide if a word belong to $L$
- typical examples: regular expression, context-free grammar
Formal language vs. Natural language

- **Formal language:**
  - concatenation of symbols to make the words of the language (possibly infinitely)
  - words have two sides: form and meaning

- **Natural language:**
  - words are concatenated to make utterances/sentences (possibly infinitely)
  - sentences have two sides: sound (or string) and meaning

→ Formalisation of grammars for Natural Language (Chomsky 1956)
Formal language vs. Natural language

But...

- Ambiguities:
  - Avoid/rejected by formal languages
  - Very important in natural languages (several linguistic structures can be associated to a sentence)
Formal language vs. Natural language

Ambiguity appears at any linguistic levels:

- phonologic: *I scream* / *Ice cream*
- lexical: (river) bank / bank (financial institution)
  unlockable
- syntactic: Mary ate a salad with spinach
  Mary ate a (salad with spinach)
  Mary (ate a salad) with spinach
Formal language vs. Natural language

- **semantic:**
  The police were ordered to stop drinking after midnight.
  A sailor was dancing with a wooden leg.
  Teacher strikes idle kids

- **anaphoric:**
  Margaret invited Susan for a visit, and she gave her a good lunch.

→ All the above sentences can be correct.

The interpretation depends on the context
How to measure effectiveness of a NLP system?

Evaluation

→ In general, evaluation measures issued from Information Retrieval and Machine Learning

Require a gold standard (all the good answers)

Let the contingency table:

<table>
<thead>
<tr>
<th>Gold standard</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td>NO</td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>
How to measure effectiveness of a NLP system?

**Evaluation**

Three measures:

- **Precision**: \( P_i = \frac{TP_i}{TP_i + FP_i} \)
- **Recall**: \( R_i = \frac{TP_i}{TP_i + FN_i} \)
- **F-measure**: avoid the difficulty to compare systems with two measures
  (harmonic mean of the precision and recall)
  \[ F_\beta = \frac{(\beta^2+1) \times P \times R}{\beta^2 P + R} \]
  (usually \( \beta = 1 \))
Gold standard

How to build a gold standard?

- Need of human annotators
- Time consuming
- Not so easy to build for some tasks
- Annotators can make different choices
  (inter-annotator agreement varies according to the task difficulties)
- Impossible to build on Terabyte of data
- Alternative: silver standard (combination of the results of several system)
Local vs. global evaluation

If there is several class, need of measuring the effectiveness at the individual or class level:

- **Microaveraging**: Sum over all the individual instances (without taking into account the class)
  - Precision: $P_\mu = \frac{\sum_{i=1}^{\lvert C \rvert} TP_i}{\sum_{i=1}^{\lvert C \rvert} (TP_i + FP_i)}$
  - Recall $R_\mu = \frac{\sum_{i=1}^{\lvert C \rvert} TP_i}{\sum_{i=1}^{\lvert C \rvert} (TP_i + FN_i)}$

- **Macroaveraging**: Evaluation by class (locally) then by averaging over the results by class (globally)
  - Precision: $P_M = \frac{\sum_{i=1}^{\lvert C \rvert} P_i}{\lvert C \rvert}$
  - Recall $R_M = \frac{\sum_{i=1}^{\lvert C \rvert} R_i}{\lvert C \rvert}$
Other evaluation metrics

- In some application there is not only one good answer (translation, rewriting, abstract definition)
  - BLEU (Bilingual Evaluation Understudy)
  - NIST metric
  - METEOR (Metric for Evaluation of Translation with Explicit ORdering)
  - ROUGE, or Recall-Oriented Understudy for Gisting Evaluation
  - ...

- The correct answer among the $n$ first ranked answers : $P@n$ (Precision among the $n$ first answer)

- Evaluation of the accuracy, sensibility, utility, ...

- How to measure the satisfaction of the final users?

Finally, comparing systems required to evaluate the statistical significance of their results (t-test, randomisation testing)
Evaluation campaigns and challenges

- Message Understanding Conferences (MUC): Information extraction (1987-1997)
- Text Retrieval TREC: Information retrieval (since 1992), Several tracks (and new tracks each years) on chemistry-related documents, medical documents, about crowdsourcing, etc.
- BioCreative (Information extraction in Biology): three campaigns since 2004
- I2B2 NLP Challenge (processing of clinical data) since 2008
- ...
How to analyse natural language data automatically?

- Need of numerous information
  - Linguistic information
  - Contextual knowledge (general and specific to the task)
- Several steps of analysis based on
  - Various approaches (formal language/grammar, rules, machine learning)
  - Linguistic resources (dictionaries)
- Evaluation based on data (no formal evaluation)
### Conclusion

Examples of full analysis of a sentence

<table>
<thead>
<tr>
<th>Types of analysis</th>
<th>The cat</th>
<th>eats</th>
<th>the mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part-of-speech</td>
<td>DET NOUN</td>
<td>VERB</td>
<td>DET NOUN</td>
</tr>
<tr>
<td>Fundamental Structure</td>
<td>Subject</td>
<td>Predicate</td>
<td></td>
</tr>
<tr>
<td>Constituents</td>
<td>SN</td>
<td>SV</td>
<td>SN</td>
</tr>
<tr>
<td>Functions</td>
<td>Subject</td>
<td>Verb</td>
<td>Object</td>
</tr>
<tr>
<td>Thematic Roles</td>
<td>Topic</td>
<td>Focus</td>
<td></td>
</tr>
<tr>
<td>Semantic Roles</td>
<td>Agent</td>
<td>Action</td>
<td></td>
</tr>
<tr>
<td>Modality</td>
<td></td>
<td>Assertion</td>
<td></td>
</tr>
</tbody>
</table>