NLP Applications

• Two main areas:
  • Massive management of textual information sources:
    • for human use
    • for automatic collection of linguistic resources
  • Person/Machine interaction
NLP Applications

- Massive management of textual information sources
  - Machine Translation
  - Information Retrieval (IR)
  - Question Answering (Q&A)
  - Information Extraction (IE)
  - Document Classification and Clustering
Machine Translation

- Process of translating a text from a source language to a target language preserving some properties
  - The main property to preserve (but not the only one) is the meaning
- MT textual vs oral
- Different degrees of human intervention
Machine Translation

- Human Translation with Machine Support
- Machine Translation with Human Support
- Fully Automated Translation
  - Empirical systems
  - Rule-based systems
  - Statistical Machine Translation
  - Example-based Translation
Some readings

- General
  - Joseba Abaitua (1997)

- SMT
  - Kevin Knight (1999)
    - [http://www.isi.edu/natural-language/people/knight.html](http://www.isi.edu/natural-language/people/knight.html)
  - Horacio Rodriguez (2001) Técnicas estadísticas para la TA
  - Cristina España (2012) Introduction to Statistical Machine Translation

Software:
- Giza++, Moses

Projects:
- MOLTO, OpenMT
• Basic approaches
  • Direct MT
  • Transfer-based
  • Interlingua-based
  • Translation Memories
• Statistic vs symbolic approaches
Machine Translation

Source text → Lexic Str. → Syntactic Str. → Semantic Str. → Interlingua

Semantic Transfer

Syntactic Transfer

Target text → Lexic Str. → Syntactic Str. → Semantic Str. → Interlingua

Direct translation

Syntactic Transfer

Semantic Transfer
## Aligned parallel corpora numbers

### Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th># segments (app.)</th>
<th># words (app.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JRC-Acquis</td>
<td>$1.0 \cdot 10^6$</td>
<td>$30 \cdot 10^6$</td>
</tr>
<tr>
<td>Europarl</td>
<td>$1.5 \cdot 10^6$</td>
<td>$45 \cdot 10^6$</td>
</tr>
<tr>
<td>United Nations</td>
<td>$3.8 \cdot 10^6$</td>
<td>$100 \cdot 10^6$</td>
</tr>
</tbody>
</table>

### Books

<table>
<thead>
<tr>
<th>Title</th>
<th># words (approx.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Bible</td>
<td>$0.8 \cdot 10^6$</td>
</tr>
<tr>
<td>The Dark Tower series</td>
<td>$1.2 \cdot 10^6$</td>
</tr>
<tr>
<td>Encyclopaedia Britannica</td>
<td>$44 \cdot 10^6$</td>
</tr>
</tbody>
</table>
Statistical Machine Translation

Mathematically:

\[ P(e|f) = \frac{P(e)P(f|e)}{P(f)} \]

\[ T(f) = \hat{e} = \arg\max_e P(e|f) = \arg\max_e P(e)P(f|e) \]
Statistical Machine Translation

\[ T(f) = \hat{e} = \text{argmax}_e P(e) P(f|e) \]

Language Model
- Takes care of fluency in the target language
- Data: corpora in the target language

Translation Model
- Lexical correspondence between languages
- Data: aligned corpora in source and target languages

\text{argmax}
- Search done by the \textit{decoder}
### Language model

\[ T(f) = \hat{e} = \arg\max_e P(e) P(f|e) \]

Estimation of how probable a sentence is.

Naive estimation on a corpus with \( N \) sentences:

Frequentist probability
of a sentence \( e \):

\[ P(e) = \frac{N_e}{N_{\text{sentences}}} \]

Problem:
- Long chains are difficult to observe in corpora.
  \[\Rightarrow\] Long sentences may have zero probability!
The n-gram approach

The language model assigns a probability $P(e)$ to a sequence of words $e \Rightarrow \{w_1, \ldots, w_m\}$.

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i | w_{i-(n-1)}, \ldots, w_{i-1})$$

- The probability of a sentence is the product of the conditional probabilities of each word $w_i$ given the previous ones.

- Independence assumption: the probability of $w_i$ is only conditioned by the $n$ previous words.
• **Translation Model** $P(f|e)$
  
  • source: $f = f_1 f_2 ... f_m$
  
  • target: $e = e_1 e_2 ... e_l$
  
  • alignment: $a = a_1 a_2 ... a_m$
  
  • in general
    - $a \in \{1,...,m\} \times \{1,...,l\}$
  
  • usually
    - $a: \{1,...,m\} \rightarrow \{0,...,l\}$
    - $a(j) \neq 0$ $f_j$ is mapped into $e_{a(j)}$
    - $a(j) = 0$ $f_j$ is not aligned

• $A(f,e)$ is the set of possible alignments ($2^{lm}$)
Statistical Machine Translation

- Translation Model $P(f|e)$.
  - One should at least model for each word in the source language:
    - Its translation,
    - the number of necessary words in the target language,
    - the position of the translation within the sentence,
    - and, besides, the number of words that need to be generated from scratch.

```
NULL Quan tornes a casa ?
```

```
When are you coming back home ?
```
Word-based models: the IBM models

They characterise $P(f|e)$ with 4 parameters: $t$, $n$, $d$, $p1$.

- **Lexical probability** $t$
  - $t(\text{Quan}|\text{When})$: the prob. that Quan translates into When.

- **Fertility** $n$
  - $n(3|\text{tornes})$: the prob. that tornes generates 3 words.

- **Distortion** $d$
  - $d(jji;m;n)$: the prob. that the word in the j position generates a word in the i position. m and n are the length of the source and target sentences.

- **Probability** $p1$
  - $p(\text{you}|\text{NULL})$: the prob. that the spurious word you is generated (from NULL).
Statistical Machine Translation

NULL Quan tornes a casa?

NULL Quan tornes tornes tornes casa?

NULL When are coming back home?

you When are coming back home?

When are you coming back home?
Expectation-Maximisation algorithm

1. Parameter initialisation
2. Alignment probability calculation
3. Parameter reestimation
4. Alignment probability recalculation
5. Converged?
   - NO: Go back to step 2
   - YES: Final parameters and alignments
Statistical Machine Translation

Decoder

\[ T(f) = \hat{e} = \arg\max_e P(e) P(f|e) \]

Responsible for the search in the space of possible translations.

Given a model (LM+TM+...), the decoder constructs the possible translations and looks for the most probable one.

In our context, one can find:

- **Greedy decoders.** Initial hypothesis (word by word translation) refined iteratively using hill-climbing heuristics.
- **Beam search decoders.**
Information Retrieval

• **Input**
  - A collection of documents
    - The Web
    - A corporate document collection
    - ...
  - A user need represented as a query

• **Output**
  - The documents of the collection that satisfy the user needs.
Information Retrieval

Queries space: Q

Documents space: D

Query

Document

Representation space: R

Human judgement: j

Representation 1

Representation 2

Comparison function: c

Oard, 1997
Ideal setting

\[ c(q(\text{query}), d(\text{doc})) = j(\text{query}, \text{doc}) \]

\[ \forall \text{query} \in Q \]

\[ \forall \text{doc} \in D \]
Information Retrieval

Textual operations

Operations over
The query

Indexing

Indexes

DB manager

Text DB

Docs retrieved

Classification

Docs classified

Searching

feedback

representation

query

User Interface

text

text

query

NLP Applications
Information Retrieval

IR types

• Type of information
  • Text, speech, structured information

• Query language
  • Exact, ambiguous

• Matching
  • Exact, approximate

• Kind of information needed
  • Loose, precise

• Relevance:
  • Usefulness of information according to user needs
Information Retrieval

- **Preprocess**
  - Lexical analysis, estandardization
    - non esstandard forms, dates, numbers, acronyms, abbreviations, idioms, ...
  - lematization
    - Morphological analysis, stemming (Porter’s stemmer)
  - filtering
  - Stopwords

- **Classification**
  - manual
  - Automatic
  - Classification vs clustering

- **Compression**
Indexing

- manual vs automatic
- indicators
  - objective: structural
  - subjective: textual (content)
- indexing pre-coordinate vs post-coordinate
- Simple terms vs Complex terms (multiwords)

Most frequent : Bag of simple words
Representing documents

- Classical Models
  - Full text
  - Boolean
  - Vectorial
  - Probabilistic

- Variants of the Probabilistic Model
  - Bayesian
  - Statistic Graphical Models

- Other paradigms
  - Generalized vectorial model
  - Extended Boolean Model
  - Latent Semantic Indexing
  - Neural Nets
**Simple Boolean Model**

Boolean expressions over terms occurring in the document (key words).
Logical connectors: AND, OR, NOT parenthesis

**Extensions**

distance constraints (at paragraph or sentence level)
Fixed or variable window

**Extended Boolean Model**

Term weighting: term frequency in the document, in the collection, normalization

**Query expansion**

- Use of external knowledge sources (e.g. WN) extension with synonyms and/or hyponyms
  - Morphological generalization
  - Relevance
  - Feedback
When the result is not a Boolean but an ordered list of documents with an associated relevance score (ranked) measures can be vectors of precision at (usually) 3, 5, 7, 9, 11 points of recall (e.g. at 0, 0.25, 0.5, 0.75, 1).
**Boolean Model**

<table>
<thead>
<tr>
<th></th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>...</th>
<th>$t_i$</th>
<th>...</th>
<th>$t_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d_2$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d_3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d_j$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**attributes**: all the terms (words, lemmas, multiwords, ...) occurring in the collection (except stopwords). Sometimes only the most frequent.

**rows**: each document represented by a vector of Booleans (1 if the term occurs in the document, 0 otherwise). For $n$ documents

**columns**: each term represented by a vector of Booleans. For $m$ terms
Vectorial Model

\[
\begin{array}{cccccc}
t_1 & t_2 & t_3 & \ldots & t_i & \ldots & t_m \\
\hline
d_1 \\
d_2 \\
d_3 \\
\vdots \\
d_j \\
\vdots \\
\end{array}
\]

\[w_{ij} \quad \text{weight (relevance) of term } j \text{ in document } i\]

**Most used way of computing relevance: TF*IDF**

\[
\begin{align*}
tf_{ij} & \quad \text{frequency of term } t_j \text{ in the document } d_i \\
df_j & \quad \text{# documents containing } t_j \\
idf_j & \quad \log (N / df_j) \\
w_{ij} & = tf_{ij} \times idf_j
\end{align*}
\]
IR and NL

• NL Resources
• NL Processors
  • Indexing
    • words, stems, lemmas, senses, multiterms
    • phrases, …
    • problems:
      • Named entities
      • Unknown words
      • Non standard units
      • polysemy
    • => Only slight improvement over using forms
• Retrieval
  • Query expansion
Cross Language Information Retrieval

CLIR, Oard, 1997

CLIR

Controlled Vocabulary

Free text

Corpus-based

Knowledge-based

parallel Corpora

comparable Corpora

Monolingual Corpora

Aligned at document level

Aligned at sentence level

Aligned at term level

Dictionary based

Ontology based

Thesaurus based
Question Answering

- Natural extension of IR
- A QA system receives a query expressed in NL and tries to provide not a document containing the answer but the proper answer (usually a fact).
- QA systems need to use NLP techniques for both processing the question and looking for the answer.
Question Answering

- Some QA systems that can be accessed through the Web:
  - START
  - IO search engine
  - Webclopedia
  - AskJeeves
    - http://www.ask.com
  - LCC
    - http://www.languagecomputer.com/
Question Answering

• Starting in TREC challenges from del TREC-8 (1999)
• Later CLEF challenges
• Related Disciplines
  • Answer Finding
    • Given a collection of questions and answers the task consists on looking for the question(s) closest to the one formulated by the user in order to provide its answer.
    • FAQ Finder: http://infolab.cs.uchicago.edu/faqfinder/
  • NL Interfaces to databases
  • Information Integration, II
  • Information Extraction, IE
  • Answer Validation Exercise (AVE)
Question Answering

- Factual QA
  - Who? When? Where?
- List QA
  - Which are the last 10 presidents of USA?
- Domain independent vs domain restricted QA
- QA with complex queries:
  - Which are the USA republican presidents after world war II?
- Linked queries
Some readings

- Horacio Rodriguez (2001)
  - http://www.lsi.upc.es/~horacio/doctorat/tapln/QA.zip
- Documentos de las conferencias TREC
  - TREC-10 http://trec.nist.gov/pubs/trec10/t10_proceedings.html

http://www.isi.edu/natural-language/projects/webclopedia/
http://www.seas.smu.edu/~sanda/
http://www.cs.utexas.edu/users/sanda/
http://www.languagecomputer.com/
http://www.dlsi.ua.es/~vicedo/
http://www.dlsi.ua.es/~antonio/
Most QA systems consist of 4 processes:

1. **Question Processing**
2. **IR of relevant documents**
3. **Segmentation in passages, IR of relevant passages**
4. **Answer Extraction**
Question Answering

Frequently performed sequentially

- Question Processing
  - Relevant terms
  - Question type
  - Focus
  - ... 

- IR of relevant documents
  - Relevant documents

- Segmentation in passages,
  - IR of relevant passages
  - Relevant passages

- Answer Extraction
  - answer
Automatic Summarization

- A summary is a reductive transformation of a source text into a summary text by extraction or generation
  - Sparck-Jones, 2001
Automatic Summarization

- Look for the relevant parts of a document and produce a summary of them
- Summarization vs IE
  - IE
    - What has to be extracted is defined a priori
      - “I am interested on this, look for it”
  - Summarization
    - An a priori definition of what is relevant is not always defined
Some readings

- **Tutorial**

- **Horacio Rodriguez (2001) Summarization**
Types of summarization

- **Type**
  - Indicative vs informative
  - Extract vs Abstract
  - Generic vs query based
  - Background vs just-the-news
  - Single-document vs multi-document
  - General vs domain restricted
  - Textual vs multimedia

- **Input**
  - Domain, genre, form, size
Automatic Summarization

- Related disciplines
  - IE, IR, Q&A, Topic identification (TI), Document Classification (DC), Event (topic) detection and tracking (TDT)
- Evaluation
- Applications
  - Biographies
  - Medical reports
  - E-mails
  - Web pages
    - Word spotters
  - News
  - Headlines extraction
    - Automatic subtitle generation
  - IR enhancements
  - Meeting interventions
Automatic Summarization

Basic schema

multi-document

single-document

query

restrictions

extract

abstract

headline
Automatic Summarization

Techniques

- Lexical chains
  - [Barzilay, 1997], [Fuentes, 2008]
- Coreference chains
  - [Baldwin, Morton, 1998]
  - [Bagga, Baldwin, 1998]
- Alignment techniques
  - [Banko et al, 1999]
- Compression, reduction or simplification of sentences (cut & paste)
  - [Jing, 2000]
  - [Jing, McKeown, 1999]
Automatic Summarization

- **Statistical models**
  - modelos estadísticos de la lengua
    - [Berger, 2001], [Berger, Mittal, 2000]
  - modelos bayesianos
    - [Kupiec et al, 1995], [Schlesinger et al, 2001]
  - cadenas ocultas de Markov
  - Regresión logística
    - [Conroy et al, 2001]

- **Machine Learning**
  - Decision trees
  - ILP
    - [Knight, Marcu, 2000], [Tzoukerman et al, 2001]

- **Similarity (and distance) measures**
  - MMR
    - [Carbonell, Goldstein, 1998]
Automatic Summarization

- **IE**
  - [Kan, McKeown, 1999]
- **Topic Detection**
  - [Hovy, Lin, 1999]
  - [Hovy, 2000]
- **Topic Signatures**
  - [Lin, Hovy, 2001]
- **Document’s rhetoric structure**
  - [Marcu, 1997]
- **Combination**
  - [Goldstein et al, 1999], [Kraaij et al, 2001],
  - [Muresan et al, 2000], [White et al, 2001].
Multidocument Summarization (MDS)

Objectives

• Summary of a collection content
• Briefing
  • concise summary of the factual matter of a set of news articles on the same or related events (SUMMONS, Radev, 1999)
• Actualization of already known information
More challenging

- Compression
- Redundancy
- Temporal terms
- Correference
Requirements

- Clustering of documents and passages
- Recall
- Anti-redundancy
- Summary cohesion
- quality
  - readable
  - relevant
  - context
- Inconsistency of sources
- Actualization
Approaches

- From the common sections of all the documents of the collection
- Common sections + unique sections
- Centroids
- Centroids + outliers
- Last document + outliers
- Common sections + unique sections + time weighting factor
Mc.Keown et al, 1999
MULTIGEN

Analysis Component

- Feature Extraction
- Feature Synthesis
- Rule Induction

Generation Component

- Theme Intersection
- Sentence Planner
- Sentence Generator

Themes

Summary

article 1 .... article n
Information Extraction

- Extracting useful information from free text
- MUC, ACE, TAC challenges
- Named Entity Recognition (NER)
- Named Entity Classification (NEC)
- Both tasks together (NERC)
- Slot Filling
- Relation Extraction
Information Extraction

NERC

y  B-PER  O  B-QNT  O  O  B-ORG  I-ORG
x  Jim  bought  300  shares of  Acme Corp.

x  B-PER  I-PER
y  Jack  London  went to  Paris

x  B-PER  I-PER
y  Paris  Hilton  went to  London
Slot Filling

- Set of relevant slots
- ML
  - Supervised Learning
  - Unsupervised Learning
    - Distant learning
  - Semisupervised Learning
    - Active Learning
- Rule-based systems
Relation Extraction

• Labeled vs unlabeled relations
• Binary vs n-ary relations
• Properties:
  • Symmetric, transitive, reflexive
• Constraints over source and target
  • NE, PER, ORG, LOC,
Relation Extraction

- ML
  - Supervised Learning
  - Unsupervised Learning
  - Semisupervised Learning
• Classification vs. Clustering
• Assign each document to one or more class(es) belonging to a predefined tagset
• Examples:
  • Spam filtering
  • Language identification
  • Level of relevance, urgency, ...
  • Thematic domain
Document Classification

• Extensions:
  • Multiclass
    • A document can be assigned to more than one class
  • Rank
    • A document is assigned to different classes according to a probabilistic distribution.

• Features
  • Textual content
  • Metadata
Document Classification

• **Approaches**
  • **Vectorial**
    • Categorize each class with a reference document (Topic Signature, Lexical Profile, ...)
    • Represent the document to classify with VSM (Vector Space Model)
    • Using a similarity measure for comparing the vector associated to the document with the reference document of each of the classes.
    • Choose the best or rank them
      • e.g. k-means
  • **ML**
    • Naive Bayes, decision lists, decision trees, maximum entropy, SVM, boosting, ...
Document Classification

- Precision = \( \frac{\text{good messages kept}}{\text{all messages kept}} \)

- Recall = \( \frac{\text{good messages kept}}{\text{all good messages}} \)