Human Language Technology: Applications to Information Access

EPFL Doctoral Course EE-724, Fall 2014
Andrei Popescu-Belis
Idiap Research Institute, Martigny

Lesson 1a: Introduction
September 18, 2014
What is human language technology?

• Information technology involving natural language, to improve productivity
  – written, spoken, or even sign language

• Another name: natural language processing
  – emphasizes theory and processing, less applications and interaction

• HLT is at the interface of several theoretical fields and has many applications
  – algorithms, machine learning, statistics
  – computational linguistics, empirical linguistics
  – human-computer interaction, interfaces
Why is HLT important?

- **For science**
  - contributes to the validation of hypotheses regarding human language, cognition, and the mind
    - test new hypotheses and theories (e.g. on language learnability)
  - important use case for information science, data processing and statistical modeling (e.g. efficiency issues)

- **For applications**
  - HLT tools bring added value to computer systems
  - as more text/speech is available online, the need for **improving access to this information** is growing
Examples of HLT applications

• Machine translation (institutionally since the 1980s)
• Spell and grammar checkers (word processors since 1990s)
• Document search (HLT+IR): local or on the Web (since 2000)
• With speech: command, instructions, dialogue, assistance
• More recently: semantic search, opinion mining, recommendation, intelligent personal assistants

• This HLT course:
  – methods to improve access to information enclosed in texts
  – overcome three barriers: quantity | cross-lingual | subjectivity
Plan of today’s lesson

1a. Introduction to HLT
   – Three barriers to text information access
   – Objectives and plan of the HLT course (fall 2014)
   – Miscellanea: prerequisites, evaluation, resources, references
   – Basic notions of computational linguistics
     • machine learning | linguistics

1b. Document classification using lexical features
   – text representation, simple classifier, experiments (TP)
I. The quantity barrier

• A lot of knowledge and information is enclosed in text-based documents
  – factual news, encyclopaedias, manuals, technical documentation, product reviews, opinions, fiction, scientific articles, etc.

• As documents become more accessible, thanks to the Internet (online publication, search engines), they also become much more numerous

➡️ The dream: make this information just as accessible as your own knowledge that is stored in your brain
  – concrete tasks: find | aggregate | discover
II. The cross-lingual barrier

• Many languages are used on the Web
  – diversity increases as users prefer their mother tongue
  – use of English as lingua franca: limited to some domains and regions (why not Chinese?)

• Translation: old problem, new solutions

➔ The dream: design software that translates automatically text or speech
  – well-defined problem, with a lot of recent progress
III. The subjectivity barrier

• IT supports more and more human interactions
  – real-time text- or speech-based (IM, Skype) or asynchronous text-based (emails, social networks etc.)
  – various forms of dialogue
  – various forms of opinion / subjectivity / polarity
    • importance of non-literal meaning
• Key information is enclosed in the interactions

➡️ The dream: decode interaction patterns to infer new knowledge, including subjective opinions
Objectives of the HLT course

• Introduce the most up-to-date methods and applications in the field: current capabilities to **overcome the three barriers**

• Demonstrate how notions from computational linguistics and machine learning can be applied to **practical tasks involving human language**

• Develop **useful skills for PhD research** in areas related to language but also in areas beyond language
  – practice of the data-driven methods (a.k.a. machine learning)
    • using large and diverse data sets
    • evaluating the performance of the resulting system
I. OVERCOMING THE QUANTITY BARRIER
   1. Document classification using lexical features
   2. Information retrieval: basics, extensions (relevance feedback; query expansion; learning to rank), recommender systems, just-in-time retrieval
   3. Question answering
   4. Graded TP: design and evaluate a simple QA system

II. OVERCOMING THE CROSS-LINGUAL BARRIER
   1. Introduction to machine translation, language modeling
   2. Translation models: phrase-based models, text alignment
   3. Decoding with phrase-based translation models
   4. MT evaluation and applications

III. OVERCOMING THE SUBJECTIVE BARRIER
   1. Detection and analysis of subjective information
   2. Content analysis of human interaction (spoken and written)
   3. Accessing the content of multimedia information (meeting browsers)

CONCLUSION: a model of HLT research, engineering and evaluation
Practical details

• Organization of the course
  – 2-hour lectures (10:15-12:00 with a break, room INM201) followed by 2-hour practical work (TP, 13:15-15:00, room CO4)
  – exercises using free software and resources
    • CO4 computers (Win 7) or your own (portability vs. platform issues)
  – feedback will be provided on all results, via email

• Grades
  – 20%: one TP report (2+2-hour work)
  – 20%: presentation of an article related to a course (15’ talk)
  – 60%: course project (3-4 days work, individually or in pairs) + report (4-8 pages) + presentation (15’ talk) with questions from the jury (expert & APB)
    • article and project to be chosen with APB, based on each participant’s interests
Practical details (continued)

• Prerequisites
  – elementary programming applied to text data processing, e.g. Perl, Java, C/C++, Python, etc.
  – statistics, machine learning, pattern recognition, AI
  – computational linguistics: not required, but helpful
  – the most important: interest & motivation

• Note: the only other EPFL course in the field is the MSc/PhD course “Introduction to NLP” by J.-C. Chappelier and M. Rajman (http://coling.epfl.ch/), given every spring
  – the two courses are complementary
Bibliography

• *Foundations of Statistical Natural Language Processing*, by Christopher D. Manning and Hinrich Schütze, MIT Press, 1999.
• *Introduction to Information Retrieval*, by Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Cambridge University Press, 2008.
• *The Handbook of Computational Linguistics and Natural Language Processing*, by Alexander Clark, Chris Fox, Shalom Lappin (eds), Blackwell/Wiley, 2010.
• *Data Mining: Practical machine learning tools and techniques*, by Ian H. Witten, Eibe Frank, and Mark A. Hall, Morgan Kaufmann, 2011.
• Archive of articles: [http://www.aclweb.org/anthology-new/](http://www.aclweb.org/anthology-new/) and use Google Scholar too
Online resources (1/2)

• Course web page (also available via the EPFL course page)
  – http://www.idiap.ch/~apbelis/hlt-course/
  – slides and TP in PDF, announcement of talks

• Software
  – WEKA toolkit for machine learning: www.cs.waikato.ac.nz/ml/weka/
  – Mallet: machine learning for language: http://mallet.cs.umass.edu/
  – many other NLP and ML packages are available
    • e.g. Stanford NLP tools, NLTK, GATE
Online resources (2/2)

- Data
  - See e.g. a list of corpora URLs on Wikipedia
  - Reuters, 20 Newsgroups, Wikipedia, etc.
  - Europarl, Hansard, JRC-Acquis
  - Online catalog of Project Gutenberg: 20,000 free e-books
  - British National Corpus, also with an online interface
- Corpus distribution agencies
  - ELDA = Evaluations and Language resources Distribution Agency
  - LDC = Linguistic Data Consortium (USA)
Machine learning and computational linguistics: a very quick tour
Machine learning for HLT

• Most HLT problems are classification problems
  ITEMS $\rightarrow$ FEATURES $\rightarrow$ CLASSES
  ‘$\rightarrow$’ is the feature extraction process
  ‘$\rightarrow$’ is the classification process (also called labeling)

• Supervised ML for classification
  – learn (= train, optimize) from already classified data
  – run (= test) the classifier on new data
    • testing: how well does it perform? – requires “new data” that is also already classified (but this is not shown to the classifier)
    • production mode: use it to label “really new data”

• Note: there are several ML courses at EPFL/EDOC
Examples of classifiers

• Classification method ≠ training method
  – *classification*: procedure to assign to an item one of the possible classes, given the values of its features
  – *training (supervised)*: procedure to set the parameters of the classifier so that it classifies correctly most of the training data
    • Note: training error is often not zero, due to inconsistencies in the data, incompleteness of features, form of the classifier, and to preserve generality

• Examples
  – decision trees (built with Id3 or C4.5)
  – KNN (no training, possibly sampling, then search)
  – Naïve Bayes (parameters estimated using frequencies, MAP decision)
  – SVM (non-probabilistic linear classifier, non-linear with a kernel)
  – neural networks (including deep NNs)
  – ensemble learning (bagging, boosting, random forests)
Training / development / testing

• Using labeled data for classification experiments
  – labeled = reference, human labels, gold standard, ground truth

• Split the data in three subsets (usually)
  1. Training set to build a classifier
  2. Development set to run tests, analyze the results, work on feature engineering or classifier selection to improve results
     – or use a single train/dev set and perform cross-validation (see next slide)
  3. Test set (held out, unseen): one final testing, report results

• Remarks
  – reporting on dev set is not informative because the system was implicitly optimized for it → how will it perform on different data?
  – the test set is also labeled, but of course the labels are not shown to the classifier, they are only used for measuring its performance
Cross-validation and significance

• With training and development data (or sometimes the entire data)
  – divide data in \( N \) folds (often 10, or 5)
  – for the \( N \) possible subsets of \( N-1 \) folds, train on each subset and test on the remaining fold
  – compute average scores and confidence intervals (related to STD) of \( N \) values, or perform paired t-tests to compare two systems
    • *significance*: what are the chances that a difference between two systems is due to the fact that one is “really better”, or that it is due to randomness?
  – easier to compute significance with c.-v. than with a single unseen test set

⇒ Training + testing = empirical or data-driven NLP
  – rigorous testing can (and must) be applied to any NLP system, even those that do not need training (e.g. hand-coded rules)
  – performance scores vary with the data set, and it is difficult to predict actual performance on a data set of a different nature than the test set
Basic linguistic concepts

• Describing human language(s) and analyzing individual linguistic productions
  – language function vs. actual languages
  – competence vs. performance

• Linguistic theories divide the description of utterances in several layers of analysis
  – sample sentence: “The little star’s beside a big star”
The little star’s beside a big star
The little star’s beside a big star
The little star's beside a big star
Layers of analysis in text vs. speech

- Letters
- Words
- Phrases
- Clause
- Sentences
- Topical units
- Texts

- Phonemes
- Syllables
- Words
- Phrases
- Clause
- Utterances
- Adjacency pairs
- Topical units
- Dialogues
Language analyses by layer

- Words: segmentation, tokenization, part-of-speech analysis, lemmatization, polarity, word sense disambiguation
- Phrases: chunking, local syntax
- Sentences: segmentation, syntax, semantic role labeling, semantic representations
- Texts: discourse relations, topics, discourse parsing
- Dialogues: dialogue acts, adjacency pairs

Available building blocks for applications, esp. lower levels
- often as free open source implementations (see slide 14)
- almost all are imperfect (well below 100%)
Conclusion

- HLT is at the cross-roads of several disciplines
- It has both scientific and practical implications
- This course will offer a perspective on recent HLT achievements and underlying methods for accessing textual information across 3 types of barriers