WHY LINGUISTICS

from a technological perspective

Toni Badia
Universitat Pompeu Fabra

content

- motivation
- Natural Language Processing
- why linguistics in NLP
- what sort of linguistics
- how should linguistic annotations be
- conclusion

motivation

- language is pervasive:
  - it plays a role
    - in understanding human behaviour
    - in interacting with one another (and building shared communication space)
    - in building knowledge, ideologies...
    - in interacting with machines
    - ...
- yet
  - the study of language is often restricted to specialised arenas
  - the field is not recognised as providing useful approaches

- internally the field is poorly structured
- there is a wide break between theoretical and applied linguistics

- NLP practitioners are quite apart from software application developers

motivation (2)

- Natural Language Processing
  - dealing with language
    - in practical contexts
    - effectively
- designing interdisciplinary EU projects
  - linguistic data need to be considered along with other sorts of data
  - the handling of linguistic data must be comparable to the rest of data
- teaching in a degree on Applied Linguistics
  - what is relevant for such a degree?
  - what approach to linguistic data is required?

Natural Language Processing

- a major shift in paradigm has occurred in the last years
  - from the 1970's
    - Artificial Intelligence and NLP were grounded on logic and linguistic knowledge
  - to now
    - success of probabilistic models (initially in speech recognition and machine translation)
    - increase in processing power and storage capacity of computers
    - confluence of NLP and corpus linguistics
  - NLP is grounded on data
    - to extract models
    - via machine learning
    - to perform the task at hand
- NLP resources are built bottom-up
  - building models from real data
  - models generalise over data
- a clear distinction is being put in practice, between
  - algorithms
  - data

data

- availability of data
  - data suited for the task
  - annotated data
    - linguistically annotated
    - with non linguistic annotations
- massive amounts of data facilitate successful NLP tasks
  - and they need not be much annotated
- but
  - very often there is no such massive data available
- then
  - smaller amounts of carefully annotated data may facilitate the task

    - » in-domain MT
    - » specific sentiment analysis
    - » detailed information extraction
    - » ...

examples of tasks

- task in opinion detection
  - in several text types: reviews, blogs, tweets
  - in several languages
  - for several topics
    - there were no data available to train classifiers
    - annotation had to combine linguistic and non-linguistic aspects
- task in sentiment analysis
  - from posts in a forum of a company
  - interest in both:
    - the opinions expressed
    - the social network structure
- task in information extraction
  - from documents in several modalities: video, audio, texts, tweets
  - in several languages
  - interactive search based on topic, named-entity...
    - to collect relevant documents (e.g., in journalistic research)
- dependency annotations
  - adapted to specific tasks

why linguistics

- linguistics is needed to make sense of linguistic facts

- linguistic facts
  - scattered
    - in multiple communications events
    - in a number of different languages
  - intersected with other communication factors and means

- linguistics is needed
  - to provide an overall picture of linguistic facts, being true to both:
    - each minute linguistic communication act, and
    - language as a cognitive capacity of human species
  - to systematically explain the correlation between form and meaning

why linguistics

- linguistic analyses are required
  - to provide data that are linguistically structured, so that
    - commonalities between different phenomena / languages emerge
      - differences in form → similar meaning
    - differences in meaning are made explicit
      - differences in form → different meaning
  - linguistic annotations must be based on sound linguistic analyses

- linguistic theories are essential
  - to provide consistency to the analyses and annotation of data

what (sort of) linguistics

- grounded on facts
- descriptive linguistics
  - descriptions that are not dependent on a specific theory
  - applicable cross-language
  - directly relating form and meaning
- validated against (quantitative) data
- truly experimental science

how should linguistic annotations be

- rigorous annotations
  - deriving from categorisation of linguistic phenomena
- using cross-lingual criteria
  - so that linguistic diversity is addressed
- quality annotations
  - simple, fully justified, grounded on facts, contrasted, replicable
- clearly quantifiable
- comparable to, and integratable with, non-linguistic data
  - image, sound, social networks... data

conclusion

- in NLP linguistics is needed
  - as far as no massive data is available for every specific task
  - to make sense of linguistic data (highly distributed)
- not all linguistics is useful
  - cross-lingual approaches are necessary
  - validation of theories / approaches must be done
    - against facts
    - taking into account the correlation between form and meaning
  - preference to rigorously annotated data
    - favouring the form – meaning mapping
  - linguistic information must be organised in a way that can be treated along with non-linguistic information
- try to answer the questions about language that non-linguists ask to linguists