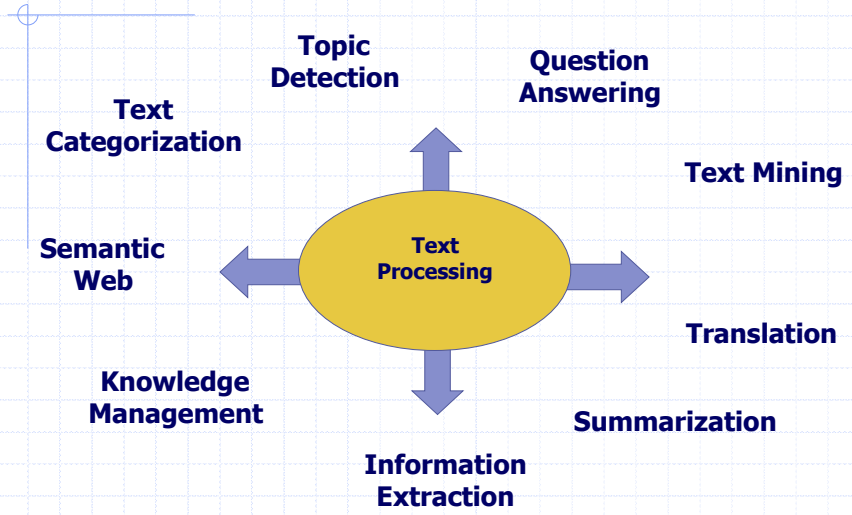
Corpora Annotations and Crouwdsourcing

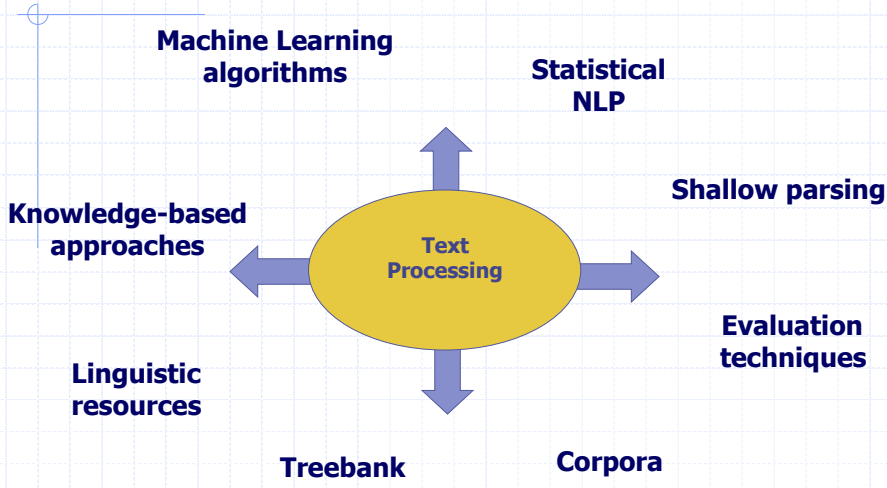
Carlo Strapparava
FBK-Irst Istituto per la ricerca scientifica e tecnologica
I-38123 Povo, Trento, ITALY
strappa@fbk.eu

Text Processing: Classical Applications



2

Text Processing: Technologies



3

The empirical turn in NLP/AI

- From systems coded by experts to large scale statistical learning from collections of "natural" texts (corpora)
- Hand-tuned systems tend to privilege "depth" over "width" and do not scale up well

The problem with linguistic experts

- Classic sentences linguists focused on:
  - Every farmer who owns a donkey beats it.
  - The idea that the idea suffices suffices.
  - The old man the boat.
- Real texts:
  - Branch Prediction Analysis is a recent attack vector against RSA public-key cryptography
  - Simon P. Chappell's review
  - (Linux.com and Slashdot are both part of OSTG.)

The empirical turn in NLP

- Focus shifts from expert system design and implementation to methods to preprocess corpora and extract information from them
- The Web is a huge database of documents, mostly text
- New problems in text processing: crawling, Web cleaning, interfacing Web services, unusual text forms

Examples of Corpora

- "Balanced"/"representative"/"reference" corpora: Brown (1M tokens), COBUILD (10M), British National Corpus (100M)
- "Opportunistic" corpora: Wall Street Jornal, Gigaword
- Parallel corpora (EuroParl)
- Manually annotated corpora for algorithm training purposes (see, e.g., the Linguistic Data Consortium catalogue)
- Corpora from/for shared competitive tasks (e.g., the SemEval or TREC corpora)

The Web as a corpus and knowledge source

- The Web is a huge database of documents, mostly text
- New problems in text processing: crawling, Web cleaning, interfacing Web services, unusual text forms
- Web-derived corpora (itWaC, deWaC, ukWac, frWaC: 1.5-2.5B tokens)
- Wikipedia as corpus
- Google trillion-word Web 1T 5-Gram collection (and Google Books Ngram collection http://books.google.com/ngrams)
- Daily patterns of life in Twitter: e.g. what the men and women talk about - http://www.tweetolife.com/gender/
- Read the Web - http://rtw.ml.cmu.edu/rtw/

Zipf's law

- If t_1 is the most common term in the collection,
  t_2 is the next most common,
  and so on,
- then the frequency freq(t_i) of the i^{th} most common term is proportional to 1/i

$$freq(t_i) \approx \frac{1}{i}$$

A Text Corpus Processing Pipeline

- Tokenization
- Part-of-speech tagging and lemmatization
- Parsing
- Word sense disambiguation
- Extraction of semantic information
- Pragmatics

The Text Processing corpus processing pipeline

A complete copy is not known to exist

A	a	DT	1	3	NMOD
complete	complete	JJ	2	3	NMOD
copy	copy	NN	3	4	SBJ
is	be	VBZ	4	0	ROOT
not	not	RB	5	4	VMOD
known	know	VVN	6	4	VC
to	to	TO	7	8	VMOD
exist	exist	VV	8	6	OBJ

What different kinds of annotation?

- *phonetic annotation*

e.g. how a word in a spoken corpus was pronounced. prosodic annotation — again in a spoken corpus — adding information about prosodic features such as stress, intonation and pauses.

What different kinds of annotation?

- *semantic annotation*

e.g. adding information about the semantic category of words — the noun *cricket* as a term for a sport or as a term for an insect belong to different semantic categories

What different kinds of annotation?

- *pragmatic annotation*

e.g. the kinds of speech act (or dialogue act) that occur in a spoken dialogue — thus the utterance *okay* on different occasions may be an acknowledgement, a request for feedback, an acceptance, or a pragmatic marker initiating a new phase of discussion.

- *stylistic annotation*

What different kinds of annotation?

- *discourse annotation*

e.g. adding information about anaphoric links in a text, for example connecting the pronoun *them* and its antecedent *the horses* in:
I'll saddle the horses and bring them round.

What different kinds of annotation?

- *lexical annotation*

adding the identity of the lemma of each word form in a text — i.e. the base form of the word, such as would occur as its headword in a dictionary (e.g. *lying* has the lemma LIE).

Useful guidelines for corpus annotation

- Annotations should be separable
  - The annotations are added as an 'optional extra' to the corpus. It should always be easy to separate the annotations from the raw corpus.
- Detailed and explicit documentation should be provided
  - How, where, when and by whom were the annotations applied?
  - What annotation scheme was applied?
  - What coding scheme was used for the annotations?
  - How good is the annotation?

Useful guidelines for corpus annotation

- *What annotation scheme was applied?*
  - An annotation scheme is an explanatory system supplying information about the annotation practices followed, and the explicit interpretation, in terms of linguistic terminology and analysis, for the annotation.
- *What coding scheme was used for the annotations?*
  - the set of symbolic conventions employed to represent the annotations themselves, as distinct from the original corpus. E.g. XML, SGML, etc.

Useful guidelines for corpus annotation

- *How good is the annotation?*
  - However, although some aspects of 'goodness' or quality elude judgment, others can be measured with a degree of objectivity.
  - Annotators should supply what information they can on the quality of the annotation.
- *How consistently has the annotation task been performed?*
  - inter-annotator agreement
  - Kappa-statistics [see Carletta, 1996]

Types of Corpus Annotation

- Part-of-speech (POS)
- Lemmatization
- Syntactical (parsing)
- Semantic (domain classifications)
- Coreference (Discourse)
- Pragmatic (Speech acts – dialogue)
- Stylistic
- Research specific (ad hoc)

POS Tagging: Claws C5

Corpus_NN1 annotation_NN1 is_VBZ
the_AT0 practice_NN1 of_PRF
adding_VVG interpretative_AJ0
linguistic_AJ0 information_NN1
to_PRP a_AT0 corpus_NN1 ._.

NN1 singular noun	AJ0 adjective (unmarked)
VBZ -s form of the verb "BE"	PRF the preposition OF
VVG -ing form of lexical verb	AT0 article

POS Tagging: POSTagger

Corpus/NN annotation/NN is/VBZ
the/DT practice/NN of/IN
adding/VBG interpretative/JJ
linguistic/JJ information/NN
to/TO a/DT corpus/NN ./.

Parsing: Chunking

[NP (NN Corpus) (NN annotation)]
(VBZ is)
[NP (DT the) (NN practice)]
(IN of) (VBG adding)
[NP (JJ interpretative) (JJ linguistic) (NN information)]
[PP (TO to) [NP (DT a) (NN corpus)]

Parsing

```
(S
   (NP Corpus annotation)
   (VP is
      (NP
         (NP the practice)
                  (PP of
               (S (VP adding
                        (NP interpretative linguistic
   information)
                        (PP to (NP a corpus))
                  ))
         )
      )
   )
   .)
```

Word Sense Annotation

- **The noun move has 5 senses (first 5 from tagged texts)**
- 1. (377) move -- (the act of deciding to do something; "he didn't make a move to help"; "his first move was to hire a lawyer")
- 2. (70) move, relocation -- (the act of changing your residence or place of business; "they say that three moves equal one fire")
- 3. (57) motion, movement, move, motility -- (a change of position that does not entail a change of location; "the reflex motion of his eyebrows revealed his surprise"; "movement is a sign of life"; "an impatient move of his hand"; "gastrointestinal motility")
- 4. (30) motion, movement, move -- (the act of changing location from one place to another; "police controlled the motion of the crowd"; "the movement of people from the farms to the cities"; "his move put him directly in my path")
- 5. (5) move -- ((game) a player's turn to take some action permitted by the rules of the game)

Word Sense Annotation

- **The verb move has 16 senses (first 13 from tagged texts)**
- 1. (130) travel, go, **move**, locomote -- (change location; move, travel, or proceed; "How fast does your new car go?"; "We travelled from Rome to Naples by bus"; "The policemen went from door to door looking for the suspect"; "The soldiers moved towards the city in an attempt to take it before night fell")
- 2. (60) **move**, displace -- (cause to move, both in a concrete and in an abstract sense; "Move those boxes into the corner, please"; "I'm moving my money to another bank"; "The director moved more responsibilities onto his new assistant")
- 3. (52) **move** -- (move so as to change position, perform a nontranslational motion; "He moved his hand slightly to the right")
- 4. (20) **move** -- (change residence, affiliation, or place of employment; "We moved from Idaho to Nebraska"; "The basketball player moved from one team to another")

An example from BNC

<s n="4"><w PRP>**After** <w AT0>**the** <w NN1>**election** <w PRF>**of** <w CRD>**28**
<w NP0>**June** <w CRD>**1973** <w AT0>**the** <w NP0>**Northern** <w NP0>**Ireland**
<w NN1>**Constitution** <w NN1>**Act** <w VVD>**became** <w NN1>**law** <w
PRP>**on** <w CRD>**18** <w NP0>**July** <w CJC>**and** <w AT0>**the**
<w NN1>**Assembly** <w VVD>**held** <w DPS>**its** <w ORD>**first** <w NN1>**meeting**
<w PRP>**on** <w AT0>**the** <w ORD>**31st**<c PUN>**.**
<s n="5"><w DT0>**This** <w VVD>**broke** <w AVP>**up** <w PRP>**in**
<w NN1>**disorder**<c PUN>**.**

An example from SemCor

<s snum=1>
<wf cmd=ignore pos=DT>**The**</wf>
<wf cmd=done rdf=group pos=NNP lemma=group wnsn=1 lexsn=1:03:00:: pn=group>**Fulton_County_Grand_Jury**</wf>
<wf cmd=done pos=VB lemma=say wnsn=1 lexsn=2:32:00::>**said**</wf>
<wf cmd=done pos=NN lemma=friday wnsn=1 lexsn=1:28:00::>**Friday**</wf>
<wf cmd=ignore pos=DT>**an**</wf>
<wf cmd=done pos=NN lemma=investigation wnsn=1 lexsn=1:09:00::>**investigation**</wf>
<wf cmd=ignore pos=IN>**of**</wf>
<wf cmd=done pos=NN lemma=atlanta wnsn=1 lexsn=1:15:00::>**Atlanta**</wf>
<wf cmd=ignore pos=POS>**'s**</wf>
<wf cmd=done pos=JJ lemma=recent wnsn=2 lexsn=5:00:00:past:00>**recent**</wf>
<wf cmd=done pos=NN lemma=primary_election wnsn=1 lexsn=1:04:00::>**primary_election**</wf>
<wf cmd=done pos=VB lemma=produce wnsn=4 lexsn=2:39:01::>**produced**</wf>
<punc>``</punc>
<wf cmd=ignore pos=DT>**no**</wf>
<wf cmd=done pos=NN lemma=evidence wnsn=1 lexsn=1:09:00::>**evidence**</wf>
<punc>''</punc>
<wf cmd=ignore pos=IN>**that**</wf>
<wf cmd=ignore pos=DT>**any**</wf>
<wf cmd=done pos=NN lemma=irregularity wnsn=1 lexsn=1:04:00::>**irregularities**</wf>
<wf cmd=done pos=VB lemma=take_place wnsn=1 lexsn=2:30:00::>**took_place**</wf>
<punc>.</punc>
</s>

CROWDSOURCING (Howe 2006)
= CROWD intelligence + outSOURCING

- "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call"

Idea:

- Collective workforce of unskilled workers equals or beats the single expert (Wikipedia-principle, many eyes see more than two)
- Micropayments for microwork unlocks workforce that cannot work a regular job
- fast throughput through potentially large number of workers

- Howe, J. (2006): "The Rise of Crowdsourcing". Wired 14.6.

Evolution of Work Organization