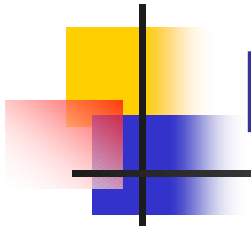
Automatic Evaluation of Robustness and Degradation in Tagging and Parsing

Johnny Bigert, Ola Knutsson, Jonas Sjöbergh

Royal Institute of Technology,

Stockholm, Sweden

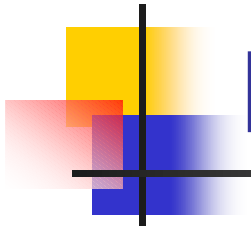
Contact: **johnny@kth.se**



Problem

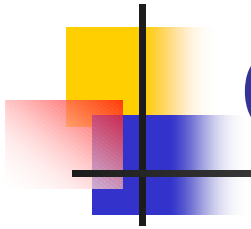
NLP systems are often faced with noisy and ill-formed input:

- How do we reliably evaluate the performance of NLP systems?
- Which methods of tagging and parsing are robust?



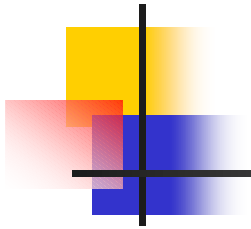
Problem

- The performance of a NLP system is sensitive to noisy and ill-formed input
- Manual evaluations of robustness is tedious and time-consuming
- Manual evaluation is difficult to compare and reproduce
- Resources with noisy data is rare



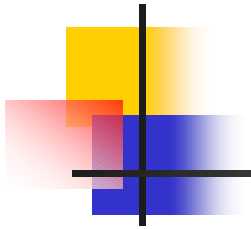
Outline

- Introduce artificial spelling errors using software (Missplel)
- Increasing error levels will affect the NLP system performance
- Evaluation of degradation of tagging and parsing performance (AutoEval)



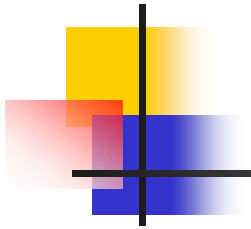
Introducing spelling errors

- Missplel (Bigert et al)
- Generic tool to introduce human-like spelling errors
- Highly configurable
- Language and tag set independent
- Freeware, open source
  http://www.nada.kth.se/theory/humanlang/tools.html



Introducing spelling errors

- Start with correct text
  (Swedish, the SUC corpus, Ejerhed et al)
- Introduce errors in, say, 10% of the words
- Spelling errors resulting in non-existing words only
- No change in parse tree



Introducing spelling errors

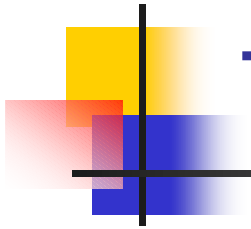
- 10 misspelled texts for each error level
- Eliminate the influence of chance

- Six error levels:
  0%, 1%, 2%, 5%, 10%, 20%
- 15 000 words with parse info

Missplel example

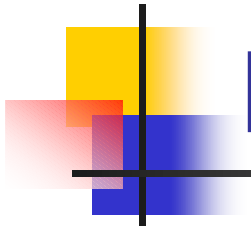
Letters	NN2
would	VM0
be	VBI
welcome	AJ0-NN1

Litters	NN2	damerau/wordexist-notagchange
would	VM0	ok
bee	NN1	sound/wordexist-tagchange
welcmoe	ERR	damerau/nowordexist-tagchange



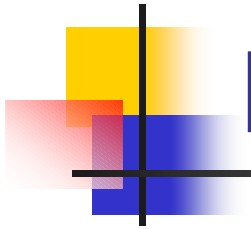
Tagging

- The texts were tagged using
  - HMM tagger (TnT, Brants)
  - Brill tagger (fnTBL, Ngai & Florian)
  - Baseline tagger (unigram)



Parsing

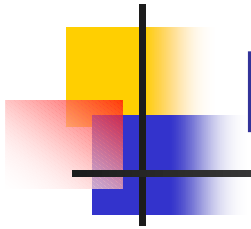
- The tagged texts were parsed using
  - GTA parser (Knutsson et al)
  - Baseline parser (unigram, CoNLL)
- GTA - Granska text analyzer
  - Rule-based
  - Hand-crafted rules
  - Context-free formalism



Parsing

Parser output in IOB format (Ramshaw & Marcus):

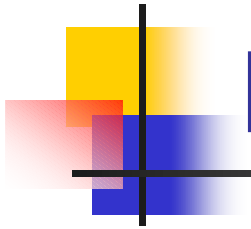
Viktigaste (the most important)	APB|NPB	CLB
redskapen (tools)	NPI	CLI
vid (in)	PPB	CLI
ympning (grafting)	NPB|PPI	CLI
är (is)	VCB	CLI
annars (normally)	ADVPB	CLI
papper (paper)	NPB|NPB	CLI
och (and)	NPI	CLI
penna (pen)	NPB|NPI	CLI
,	0	CLB
menade (meant)	VCB	CLI
han (he)	NPB	CLI
.	0	CLI



Evaluation

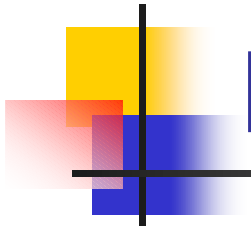
Evaluation was carried out using AutoEval (Bigert et al):

- Automated handling of plain-text and XML input/output and data storage
- Script language
- Highly configurable and extendible (C++)
- Freeware, open source **http://www.nada.kth.se/theory/humanlang/tools.html**



Evaluation

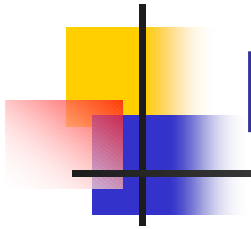
- Tagging:
  - Accuracy, correct tag if exact match
- Parsing:
  - Accuracy, correct row if exact match
  - Precision and recall per phrase category, correct if exact match after removing all other phrase types
- Clause boundary identification
  - Precision and recall for CLB



Results

Results of the tagging task (accuracy):

Tagger	**0%**	**1%**	**2%**	**5%**	**10%**	**20%**
Base	85.2	84.4 (0.9)	83.5 (1.9)	81.2 (4.6)	77.1 (9.5)	69.0 (19.0)
Brill	94.5	93.8 (0.7)	93.0 (1.5)	90.9 (3.8)	87.4 (7.5)	80.1 (15.2)
TnT	95.5	95.0 (0.5)	94.3 (1.2)	92.4 (3.2)	89.5 (6.2)	83.3 (12.7)

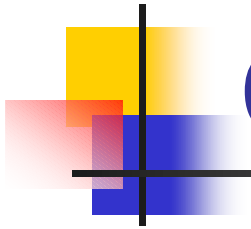


Results

Results of the parsing task (accuracy):

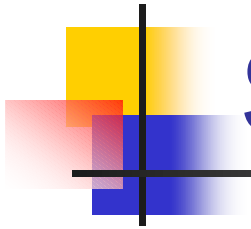
Tagger	**0%**	**1%**	**2%**	**5%**	**10%**	**20%**
Base	81.0	80.2 (0.9)	79.1 (2.3)	76.5 (5.5)	72.4 (10.6)	64.5 (20.3)
Brill	86.2	85.4 (0.9)	84.5 (1.9)	82.0 (4.8)	78.0 (9.5)	70.3 (18.4)
TnT	88.7	88.0 (0.7)	87.2 (1.6)	85.2 (3.9)	81.7 (7.8)	75.1 (15.3)

Baseline parser: 59.2% at the 0% error level, using TnT



Conclusions

- Automated method to determine the robustness of tagging and parsing under the influence of noisy input
- No manual intervention
- Greatly simplifies repeated testing of NLP components
- Freeware



Software

- Missplel and AutoEval
- Open source
- Available for download at
  the Missplel and AutoEval homepage

  `http://www.nada.kth.se/theory/humanlang/tools.html`