

Evaluating Natural Language Processing Tasks with Low Inter-Annotator Agreement: The Case of Corpus Applications

Vojtěch Kovář

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno
`xkovar3@fi.muni.cz`

RASLAN 2016



Partially supported by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.



Outline

- 1 Goal of NLP
- 2 Gold standards
- 3 What's Wrong?
- 4 Solution?
- 5 Applications
- 6 Conclusions

Goal of Natural Language Processing

■ Applications

- to (help us) translate a text
- to summarize text for us
- to answer our questions
- ...

■ Is it a trivial fact?

- most scientific papers do not evaluate applications
- but they should

Gold Standards

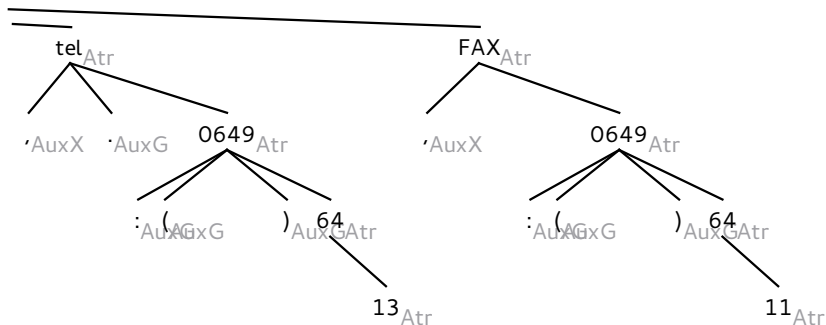
- State-of-the-art methodology for evaluating NLP tasks
- Gold standard
 - a data set manually annotated for correct solutions
 - syntactic analysis: treebanks
 - machine translation: parallel corpora
 - corpora with annotated named entities
 - ...
- Evaluation
 - comparing output of a tool with the gold standard
 - precision & recall, similarity scores

What's Wrong?

■ Overfitting to gold standard

- creating gold standards is expensive
- (unlike the evaluations themselves)
- \Rightarrow one (or a few) gold standards per task
- \Rightarrow one type of output for all tools, defined by gold standard
- \Rightarrow e.g. same granularity
- **this does not correspond to the reality**
- different applications need different information
- e.g. recognizing named entities in Wikipedia vs. on Facebook

What's Wrong? – an Example



What's Wrong? – II

- Inter-annotator agreement in gold standards

- often low
- rarely published

- Often unreachable

- syntactic analysis: $< 95\%$
- terminology extraction?
- topic recognition?
- text summarization?
- **but we need to evaluate these as well**

- Trying to increase the agreement

- extensive manuals (300 pages for Penn Treebank)
- arbitrary decisions rather than language understanding

What's Wrong? – III

■ Arbitrary decisions are crucial

- comparing syntactic parsers for Czech on two different gold standards
- → **negative correlation**
- which one is correct?

■ Arbitrary decisions do not correspond to application needs

- **negative correlations** between gold standard evaluations and application-based evaluation
- or weak improvements, compared to gold standard evaluations
- (various syntactic analysis tasks)

Solution?

- Evaluate final applications only

- rather than measure similarities to what we **expect** to be useful
- because the gold standard designers do not really know what the application needs

- This means

- find/create application that will benefit from the tool
- evaluate results of the application

- In some cases, this means asking people

- more expensive & subjective
- less replicable & sensitive
- **but** the only way to measure what we really need

Discussion

■ Price

- evaluation by humans is more expensive
- but it is not needed too often
- in development: automated tests for regression

■ Replicability & sensitivity

- human evaluations are not perfectly replicable
- (automated evaluations *sometimes* are)
- but they are replicable **to a significant extent**
- measuring 0.1% differences does not make sense anyway

Discussion – II

■ Specificity

- the results will be application-specific
- **but** that is much better than results irrelevant to any application

■ Subjectivity, more space for cheating

- it is possible to cheat with gold standards, too
- human evaluations technically easier to replicate/disprove

Evaluation of Corpus Applications

- Word sketch (collocation extraction)
- Distributional thesaurus
- Terminology extraction
- Principle
 - select a suitable sample
 - show 2 versions of the output to the evaluators
 - let the evaluators judge parts where they differ
 - (and nothing else)
 - sum the judgements

buy ^(verb)
British National Corpus (BNC)

buy ^(verb) English Web 2008 (enTenTen08)

<u>modifiers of "buy"</u>	<u>+</u>	<u>-</u>
<u>2,775</u>	<u>0.11</u>	
cheaply	<u>16</u>	7.45
bought cheaply in		
recently	<u>58</u>	7.07
recently bought a		
privately	<u>14</u>	6.92
separately	<u>13</u>	6.74
be bought separately .		
some	<u>11</u>	6.64
bought some		

<u>modifier</u>	<u>+</u>	<u>-</u>
<u>74,198</u>	<u>0.13</u>	
recently	<u>1,569</u>	6.91
locally	<u>299</u>	6.56
just	<u>6,715</u>	6.44
actually	<u>1,683</u>	6.16
dearly	<u>168</u>	6.09

<u>objects of "buy"</u>	<u>+</u>	<u>-</u>
<u>13,114</u>	<u>0.53</u>	
house	<u>506</u>	9.40
share	<u>271</u>	8.98
buy shares		
ticket	<u>219</u>	8.93
car	<u>264</u>	8.60
good	<u>190</u>	<u>+</u> <u>-</u>

<u>object</u>	<u>+</u>	<u>-</u>
<u>329,714</u>	<u>0.59</u>	
ticket	<u>8,838</u>	8.87
car	<u>7,089</u>	7.43
house	<u>7,841</u>	7.30
CD	<u>2,040</u>	<u>+</u> <u>-</u>
share	<u>3,353</u>	6.88

<u>subjects of "buy"</u>	<u>+</u>	<u>-</u>
<u>3,381</u>	<u>0.14</u>	
customer	<u>47</u>	7.97
customers buy		
investor	<u>27</u>	7.61
consumer	<u>25</u>	7.59
dealer	<u>23</u>	<u>+</u> <u>-</u>
collector	<u>18</u>	<u>+</u> <u>-</u>

<u>subject</u>	<u>+</u>	<u>-</u>
<u>78,061</u>	<u>0.14</u>	
investor	<u>663</u>	6.40
consumer	<u>877</u>	5.97
i	<u>1,870</u>	<u>+</u> <u>-</u>
n't	<u>1,531</u>	<u>+</u> <u>-</u>
customer	<u>1,031</u>	5.51

<u>"buy" and/or</u>	<u>+</u>	<u>-</u>
<u>1,073</u>	<u>0.04</u>	
sell	<u>433</u>	12.62
buying and selling		
rent	<u>35</u>	9.97
buy or rent		
go	<u>118</u>	8.88
go and buy		
hire	<u>16</u>	8.79
buy or hire		

<u>and/or</u>	<u>+</u>	<u>-</u>
<u>26,672</u>	<u>0.05</u>	
sell	<u>10,916</u>	9.07
rent	<u>813</u>	8.32
lease	<u>265</u>	7.23
borrow	<u>184</u>	5.88
resell	<u>55</u>	5.83

buy (verb)

English Web 2008 (enTenTen08)

<u>object</u>			<u>objects of "buy"</u>		
	<u>329714</u>	0.59		<u>4658048</u>	0.59
ticket	<u>8838</u>	8.87	viagra	<u>119574</u>	 
car	<u>7089</u>	7.43	buy viagra		
house	<u>7841</u>	7.3	house	<u>97240</u>	8.98
CD	<u>2040</u>	 	buy a house		
share	<u>3353</u>	 	ticket	<u>76829</u>	8.92
			car	<u>94024</u>	8.9
			product	<u>111251</u>	 

buy

English Web 2013 (enTenTen13)

<u>subject</u>			<u>subjects of "buy"</u>		
	<u>78061</u>	0.14		<u>979332</u>	0.12
investor	<u>663</u>	6.4	viagra	<u>19595</u>	9.11
consumer	<u>877</u>	5.97	viagra buy		
i	<u>1870</u>	5.74	ciali	<u>14650</u>	8.83
n't	<u>1531</u>	5.57	cialis buy		
customer	<u>1031</u>	5.51	store	<u>12824</u>	8.08
			store bought		
			i	<u>60675</u>	8.0
			where can i buy		
			customer	<u>16616</u>	7.86

<u>modifier</u>			<u>modifiers of "buy"</u>		
	<u>74198</u>	0.13		<u>929179</u>	0.12
recently	<u>1569</u>	6.91	recently	<u>15410</u>	7.0
locally	<u>299</u>	6.56	recently bought a		
just	<u>6715</u>	6.44	cheap	<u>3650</u>	 
actually	<u>1683</u>	 	buy cheap		
dearly	<u>168</u>	 	locally	<u>4228</u>	6.86
			buy locally		
			just	<u>83804</u>	6.83
			just bought		



























<u>and/or</u>			<u>"buy" and/or ...</u>		
	<u>26672</u>	0.05		<u>327111</u>	
sell	<u>10916</u>	9.07	sell	<u>133683</u>	
rent	<u>813</u>	8.32	buying and selling		
lease	<u>265</u>	7.23	rent	<u>10837</u>	
borrow	<u>184</u>	 	buy or rent		
resell	<u>55</u>	 	go		 
			go and buy		
			lease		 
			buy or lease		



























phone ^(noun)
English Web 2008

Lemma		Score	Freq
telephone		0.423	124,738
computer		0.410	530,883
device		0.372	296,167
camera		0.365	209,115
PC		0.336	134,399
card		0.334	349,935
radio		0.328	200,515
machine		0.321	278,444
TV		0.321	265,480
laptop		0.303	43,108

phone ^(noun)
English Web 2013

Lemma		Score	Freq
device		0.555	3,304,299
computer		0.537	3,905,281
card		0.478	5,167,248
internet		0.469	4,614,065
iphone		0.463	1,152,330
machine		0.461	2,561,755
camera		0.457	2,147,621
app		0.450	2,028,069
website		0.450	6,863,922
network		0.438	3,914,201

Single-word	Multi-word
<input type="checkbox"/> co2  	<input type="checkbox"/> climate change
<input type="checkbox"/> biodiversity  	<input type="checkbox"/> greenhouse gas
<input type="checkbox"/> ecosystems	<input type="checkbox"/> water quality
<input type="checkbox"/> emissions	<input type="checkbox"/> carbon dioxide  
<input type="checkbox"/> unep  	<input type="checkbox"/> renewable energy
<input type="checkbox"/> watershed  	<input type="checkbox"/> sea ice  
<input type="checkbox"/> deforestation  	<input type="checkbox"/> global warming  
<input type="checkbox"/> climate  	<input type="checkbox"/> global climate  
<input type="checkbox"/> biomass  	<input type="checkbox"/> fossil fuel
<input type="checkbox"/> habitats  	<input type="checkbox"/> sustainable development  

Single-word	Multi-word
<input type="checkbox"/> sustainability  	<input type="checkbox"/> renewable energy
<input type="checkbox"/> ecosystem  	<input type="checkbox"/> climate change
<input type="checkbox"/> solar  	<input type="checkbox"/> greenhouse gas
<input type="checkbox"/> epa  	<input type="checkbox"/> clean energy  
<input type="checkbox"/> renewable  	<input type="checkbox"/> energy efficiency  
<input type="checkbox"/> ecosystems	<input type="checkbox"/> solar power  
<input type="checkbox"/> wetlands  	<input type="checkbox"/> water quality
<input type="checkbox"/> pv  	<input type="checkbox"/> solar energy  
<input type="checkbox"/> sustainable  	<input type="checkbox"/> food security  
<input type="checkbox"/> emissions	<input type="checkbox"/> fossil fuel

Conclusions

- There are problems in gold standard evaluation methodology
 - which is currently almost a dogma
 - and used rather mechanically
- Final applications should be taken into account in evaluations
 - we propose to use **only** evaluations based on applications
- It is a question of “evaluation culture” in NLP
 - let's change it!

We have introduced an evaluation scenario for 3 corpus practical applications with low inter-annotator agreement.