AGREE: A new dataset for evaluating language models

Vít Baisa

Natural Language Processing Centre
Faculty of Informatics
Masaryk University

Karlova studánka, RASLAN, 2016

Motivation

- perplexity, entropy used for LM evaluation
- English data sets prevalent
- easier interpretation of LM performance
- inspired by Microsoft Sentence Completion Challenge
- exploiting our large web-based corpora

Microsoft Research Sentence Completion Challenge

- data set published by Microsoft Research team
- training data: 522 19th-century novels (200 MB, from Project Gutenberg)
- evaluation data: 1,000 sentences from 5 Sherlock Holmes novels
- one missing word in each sentence
- the task is to choose one of 5 possible words
- LMs assign probabilities to the 5 sentences, the best one is chosen

Examples

The stage lost a fine _____ , even as science lost an acute reasoner, when he became a specialist in crime.

a) linguist b) hunter c) actor d) estate e) horseman

What passion of hatred can it be which leads a man to _____ in such a place at such a time.

a) lurk b) dine c) luxuriate d) grow e) wiggle

My heart is already _____ since I have confided my trouble to you.

a) falling b) distressed c) soaring d) lightened e) punished

Czech grammar agreement task

- choosing the proper verb suffixes in past tense, 5 options
- _, a, o, y, i
- dělal, dělala, dělalo, dělaly, dělali
- determined by grammar agreement between subject and predicate: gender, number, person, animacy
- semantics influence the suffixes
- sometimes even common sense is needed (unexpressed subject)
- sometimes not possible to determine, but possible to exclude

Data set

- czTenTen corpus used
- only sentences containing past tense verbs
- 40–120 character long
- starting with uppercase letters
- TRAIN 9,900,000 sentences
- VALID 99,000 sentences
- EVAL 996 sentences

Examples

Pestrý program *byl*_ vítanou inspirací pro naše soubory.

Určitě tady všichni nešťastnému dědulovi *držel*_ palce, ale to *byl*_ asi všechno, co pro něho *mohl*_ udělat

Léon Bourgeois navrhl_ i praktický program solidarity.

Teď už se normálně postavil_ a ťapkal_ trávou směrem ke mně.

Ve třetím kole narazil_ na celkově třetí Třešňákovou s Pilátovou a podlehl jim 0 : 2 (-18, -8).

The task

Challenging for language models since

- Czech is morphologically rich
- free word order
- high OOV rate
- predicate and subject far from each other
- semantically driven agreement

Language models on AGREE

Model	Accuracy
Human (average)	86.5
Recurrent Neural Network with hidden layer 100	59.8
SRILM word-based 4-gram	59.6
Chunk-based language model	58.7
SRILM character-based 9-gram	53.9
Baseline (the most frequent wordform)	42.0
Random (average of 10 runs)	19.6

Conclusion, future work

- data was released with scripts
- `nlp.fi.muni.cz/~xbaisa/agree`
- part of a suite for LM evaluation?
- served for argumentation in my PhD thesis

- Estonian cases
- diacritics restoration (Greek, Romanian, Vietnamese, Igbo)