Web crawling, Discrimination of similar languages, Active learning & Genre classification

Vít Suchomel

Seminář zpracování přirozeného jazyka

12. října 2015

Web crawling, SpiderLing issues

- The crawler deadlocks after some time of big scale crawling – still searching for the cause.
- Non-critical performance bottlenecks optimised.
- Crawling in 2016: Italian (5.67 G words), Romanian (2.64 G words), Belarusian (in progress), Czech (in progress).

Discrimination of similar languages

- DSL competition in a COLING workshop
- Languages (tasks A, B): sr/hr/bs, fr-CA/fr-FR, es-AR/es-ES/es-MX, pt-BR/pt-PT, my/id
- Languages (task C): Arabic dialects – EGY, GLF, LAV, MSA, NOR
- Team: "Perfect is the enemy of good": Vít S. + Vít B. + Ondra H. + Pary
- Ondra H. implemented Pary's expectation–maximisation algorithm
- Vít B. used his Chunk based language model (1–256 byte chunks, at least 3 occurrences in training data)

Pary's E–M approach in a nutshell (I)

- Given a sentence of words, the aim is to find the language with the highest probability "the sentence is in the language".
- That can be reduced to maximizing probabilities of separate words belonging to the language.
- The iterative algorithm is initialised with relative counts of words in texts in the language from the training set or from big web corpora.
- Some words occur quite frequently in a single language (and significantly infrequently in other languages) while other words occur in multiple languages.

Pary's E–M approach in a nutshell (II)

- $P(word|language)$ is represented by a function that is calculated for each word and language (the E step).
- The relevancy (weight) of words for a language is represented by "latent variable" λ for each word. The variables are updated in each iteration to give more weight to words from more relevant sentences with regards to the language (the M step).

$$P'(w|lang) = \frac{\sum_{sent}^{sentences} \lambda_{lang}^{\alpha}(sent) \cdot \frac{\lambda_{lang}(sent) \cdot P(w|lang) \cdot \mathrm{count}_{sent}(w)}{\sum_{lang}^{languages} \lambda_{lang}(sent) \cdot P(w|lang)}}{\sum_{sent}^{sentences} \lambda_{lang}^{\alpha}(sent) \cdot \lambda_{lang}(sent) \cdot |sent|}$$

Guess the language – Task A & Results

Sentences of 5+ words ... some garbage

- Međutim, tu ga je čekalo ne baš prijatno iznenađenje. Jedan obožavatelj je ushićeno počeo da viče. [sr]
- 6. J.-M. Latvala (Ford Focus) a 1.06,4 [es-ES]
- 19. Christina Staudinger (AUT) 1'44"02 [fr-FR]

Closed Run	**Accuracy**	**F1**
EM, 0 iterations	0.8651	0.8643
EM, 1 iteration	0.8659	0.865
CBLM	**0.8827**	**0.8829**

Guess the language – Task B & Results

Short sentences from social networks ... much garbage

- RT @ZarskeMauricio: Não se ilude não, se a pessoa ta online e não ta falando contigo é pq tem alguém mais interessante que você, falando co... [pt-BR]
- Nasao sam ljubav na Trebevicu <U+1F602> <U+1F604> #selfiemaniac #nosikiriki #sarajevo @Napretkov dom "Trebević" https://t.co/2h3VawuD37 [bs]
- omggg https://t.co/cL8Xe9lAOM [pt-BR]

Closed Run	**Accuracy**	**F1**
E–M, 0 iterations	**0.8**	**0.7929**
E–M, 1 iteration	0.712	0.7392
CBLM	0.424	0.4557

Open Run	**Accuracy**	**F1**
E–M, 0 iterations	**0.8**	**0.8149**
E–M, 1 iteration	0.51	0.6484

Guess the language – Task C & Results

Sentences of 1+ words, Arabic dialects . . . hard

- bAb HrmwA AltElym wHrmwA AlvqAfp wslymh kAfp AlHqwq lm yEd >mAmhm <lA >n ynthy mhnp bAstErAD Alqwp >w bsT AlsyTrp wAlnfw* fwAjb wtjAwbA mE slTp jdydp tHAwl >n tjd >dwAt Alty tsyTr bhA fhy lA tstTyE >n tsyTr bAlsyAsp wlA bAlvqAfp wlA b>y mnTq w<nmA tsyTr bAlbTl [MSA]
- Ark [EGY] Ark [GLF] Ark [LAV] Ark [NOR]

Closed Run	**Accuracy**	**F1**
E–M, 0 iterations	0.3961	0.3666
E–M, 1 iteration	**0.461**	**0.4516**
CBLM	0.4474	0.4473

Error analysis – EM-0 vs. CBLM

EM-0 right, CBLM wrong

- 9. (9) Juan Martin Del Potro (Arg) 3180 [es-AR, CBLM: hr]
- * Sainte-Catherine 3,65$ * Mont-Tremblant 4,13$ * Blainville 25,52$ * Drummondville 106,63$ * Farnham 183,02$ [fr-CA, CBLM: id]
- 2) O PAÍS - p. 4 - Para acalmar aliados, Dilma afaga ministros [pt-BR, CBLM: es-MX]

EM-0 wrong, CBLM right

- Protagonizada por: Brad Pitt, Jonah Hill, Philip Seymour Hoffman, Chris (I) Pratt. [es-MX, EM-0: bs]
- 23. Charles Pic (França), Marussia-Cosworth, 1m42.675s [pt-PT, EM-0: fr-FR]
- Martes a Sábados 8 a 20/Domingos 9 a 13 [es-AR, EM-0: pt-PT]
- Bugalhos . . . . . 1009 0 0 [pt-PT, EM-0: my]
- — (((— Israel — 2007 — 85 minutos — Género-Comedia — Director-Eran Kolirin — Actores-Saleh Bakri, Ronit Elkabetz, Sasson Gabai — [es-MX, EM-0: my]

Competition vs. real world deployment

Competition

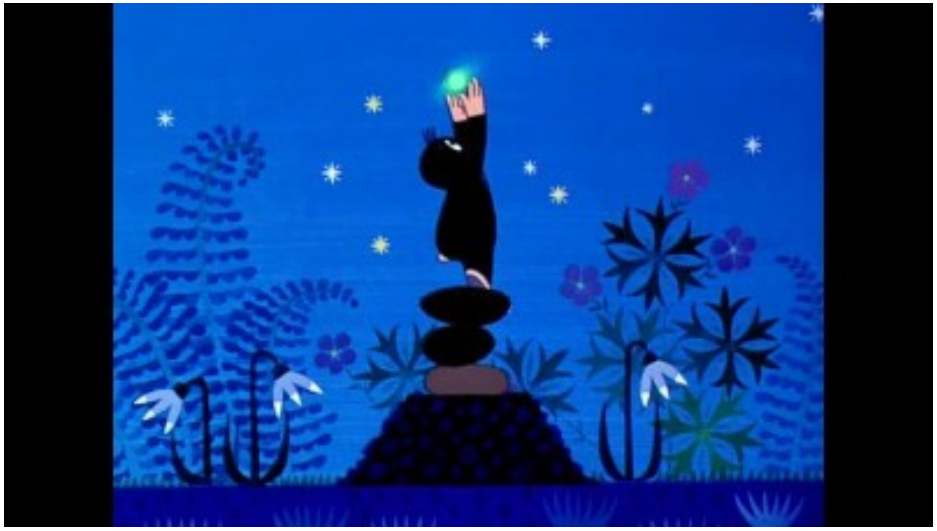
- Training data = sentences, short sentences, socnet messages ⇒ some instances hardly usable
- Balanced classes ⇒ good for machine learning
- Balanced classes ⇒ move unsure samples from larger class to a smaller one (?)
- Letting CBLM decide where EM is unsure would improve the result (?)

Web text processing for corpus linguistics

- Documents, paragraphs, sentences
- Discriminating languages in unbalanced texts (filtering out Danish and Swedish from Norwegian)
- Separating a minority dialect represented by less text from a standard text in a language (Norwegian written standards: Bokmål (85–90 %) and Nynorsk)
- Joining both methods might be impractical

Active learning

- High cost of annotation, many annotations needed, samples easy to obtain
- Active learning: Select the most useful unlabelled instances for manual annotation, re-train, iterate
- Uncertainty sampling: the most useful samples = the most uncertain = those closest to the threshold (binary classification with a threshold).



Sample selection process

Stay in Leeds with prof. Sharoff

- Use supervised learning to optimise a threshold for binary classification of a multidimensional problem
- Use active learning and measure uncertainty
- Apply genre classification works to large web corpora
- Get to know more about the content of web corpora in Sketch Engine