Spellchecker for Ukrainian

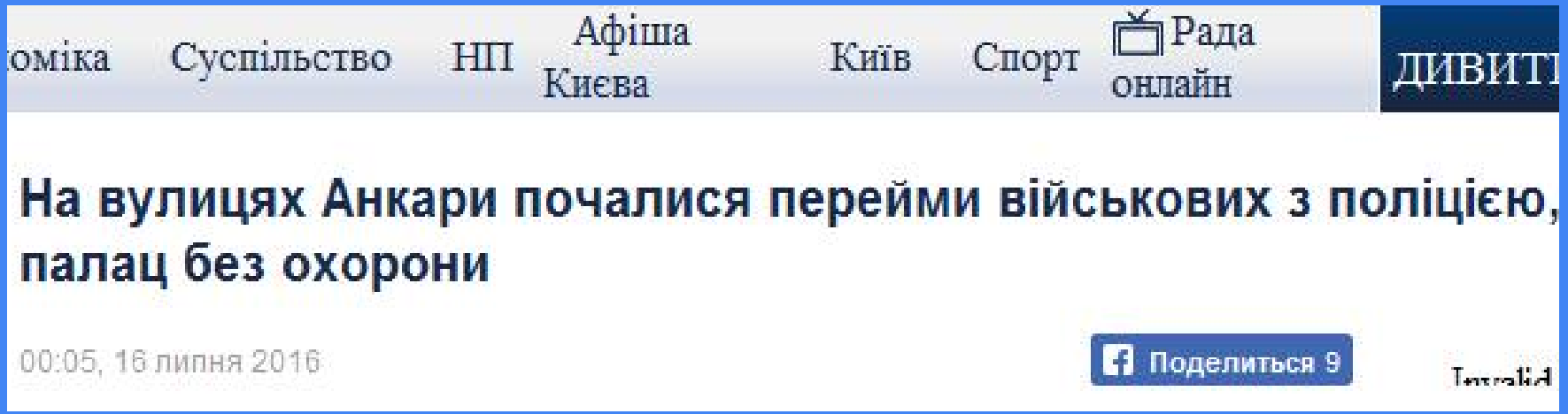
aka Spellchuk

Participants: Natalia Cheilytko, Mykhailo Kotov, Vasyl Starko

Mentor: Jordi Carrera Ventura

Data Science Summer School @ UCU
Lviv 2016

Machine-translated text



The military and the police went into labor pains in the streets of Ankara ???

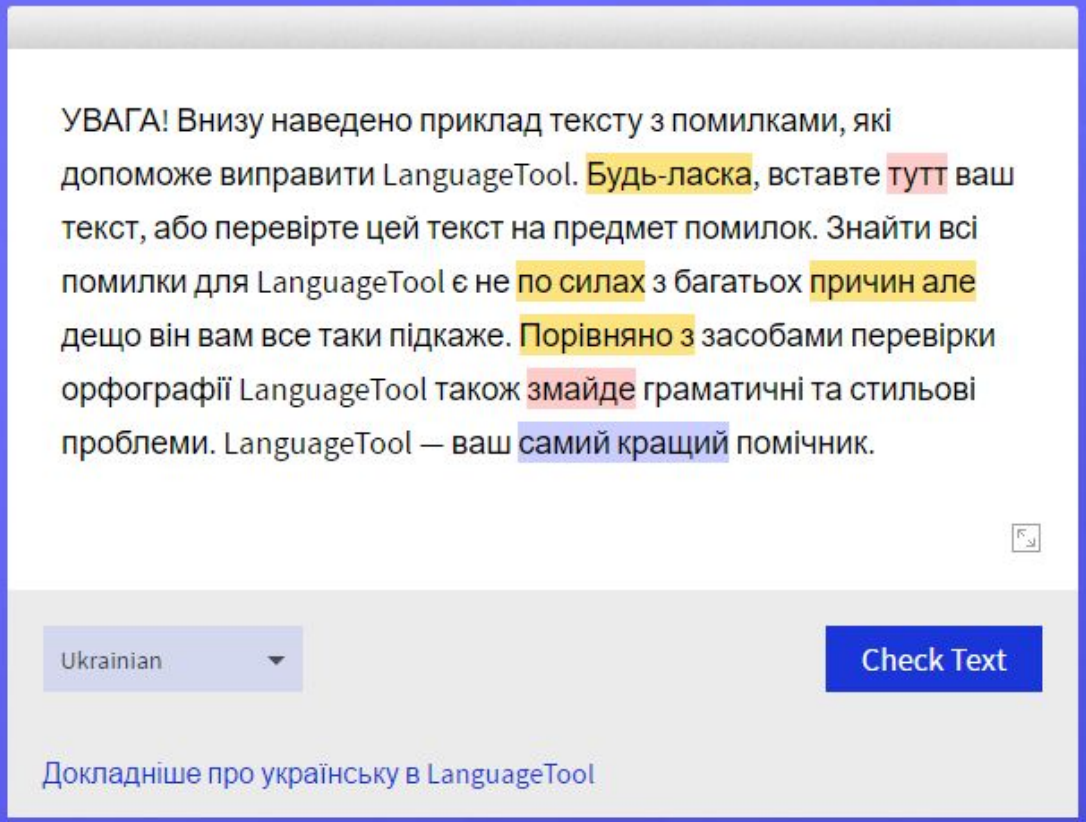
Disambiguation

Великі дані vs. дані проекти

За даними уряду vs. за даними фактами

База даних vs. у даних умовах

Foundational project



LanguageTool is an Open Source proof-reading program for English, French, German, Polish, and more than → 20 other languages.

It finds many errors that a simple spell checker cannot detect and several grammar problems.

languagetool.org/uk

Kaggle Tutorial

Completed • Knowledge • 578 teams

Bag of Words Meets Bags of Popcorn

Tue 9 Dec 2014 – Tue 30 Jun 2015 (13 months ago)

https://www.kaggle.com/c/word2vec-nlp-tutorial

Workflow

- Data acquisition and preparation
- Training Word2Vec for Ukrainian
- Building 2 Random Forest classifiers to:
  - solve the ***дані*** case
  - detect unedited machine-translated texts

...from scratch

Fieldwork

WordWithError	Correction	ErrorCategory	Machine Translation Feature	Context	Source (URL for Russian Text)
півтора роки	півтора року	grammar		Обох зловмисників засуджено до позбавлення волі терміном на півтора роки.	http://www.pravda.com.ua/news/2016/06/1/7110464/
контрсанкций	контрсанкцій	spelling	Y	Воно також визначить і допустимі обсяги ввезення виведених з-під контрсанкций продуктів.	http://www.pravda.com.ua/news/2016/06/1/7110451/
зафлудили	заполонили	lexical		Кремлівські опоненти також звинувачують онлайн "тролів", яких фінансує російська влада, які зафлудили Twitter і Facebook скаргами, щоб задушити критику Москви в Інтернеті.	http://www.pravda.com.ua/news/2016/06/1/7110400/
зі звинувачення	зі звинуваченням	grammar	Y	Кремль же неодноразово виступає зі звинувачення, що західні ЗМІ та іноземні уряди намагаються очорнити Путіна і його владу.	http://www.pravda.com.ua/news/2016/06/1/7110400/
на території	на території	spelling	Y	Для розуміння, сьогодні ні органи місцевого самоврядування, ні Держекоінспекція не можуть назвати точної кількості сміттєзвалищ, які є на території країни.	http://www.pravda.com.ua/news/2016/06/1/7110352/
благоустрій	доброустрій, порядок, добрий лад	lexical	Y	Штрафи за викинуте сміття й порушений благоустрій можуть зрости в кілька разів – до 17 тисяч гривень.	http://kiev.pravda.com.ua/news/5751a8571e215/
більше місяця	понад місяць	grammar		Як з'ясувалося згодом, таких відмов вболівальникам є чимало, а дехто з них змушений чекати на рішення консула більше місяця.	http://www.eurointegration.com.ua/news/2016/06/3/7050281/
обгрунтувати	обґрунтувати	spelling	Y	Попри надані посольству Франції документи, консул ухвалив рішення, що "інформація, надана, щоб обгрунтувати цілі та умови можливого перебування, не виглядає переконливою".	http://www.eurointegration.com.ua/news/2016/06/3/7050281/
Солошенка і Афанасьєва	Солошенка й Афанасьєва	combinatory	Y	Президент "обережно прогнозує" звільнення Солошенка і Афанасьєва	http://www.pravda.com.ua/news/2016/06/3/7110683/
Відносно прогресу	стосовно/ щодо прогресу	lexical		Відносно прогресу в справі Солошенка і Афанасьєва президент зазначив: "Не виключаю, що це може відбутися в червні".	http://www.pravda.com.ua/news/2016/06/3/7110683/

Word2Vec Parameters

Architecture: Skip-gram (default). Slower but produced better results

Training algorithm: Hierarchical softmax (default)

Downsampling of frequent words: 0.001

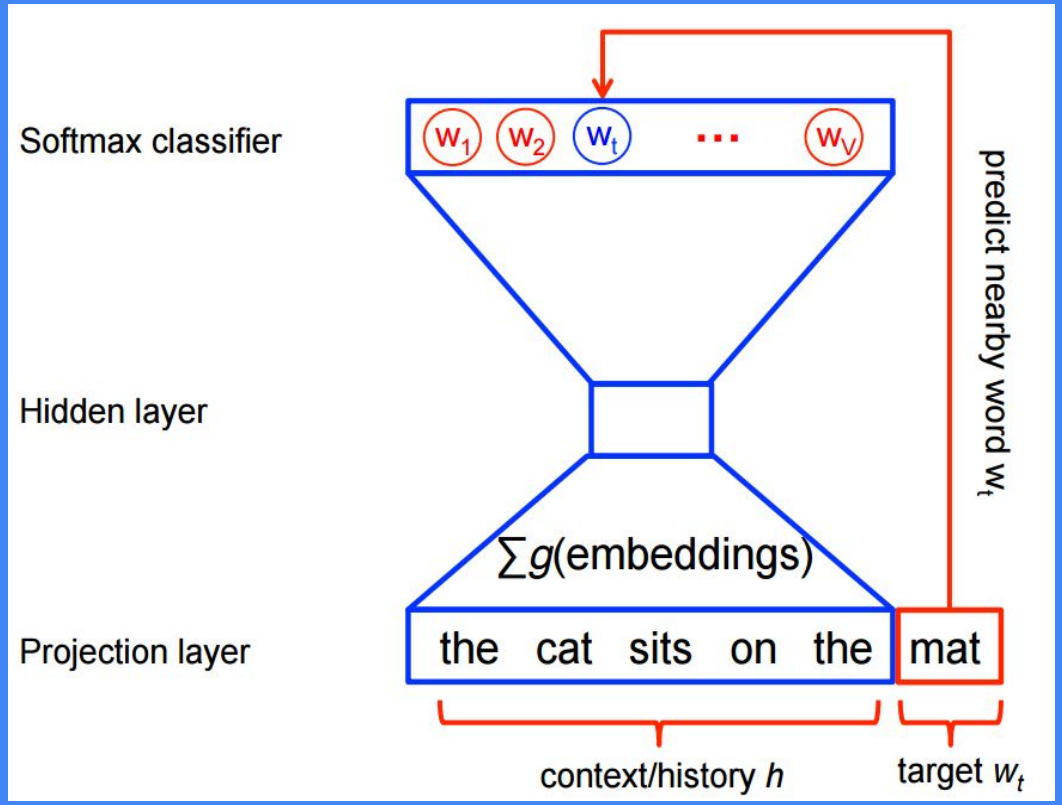
Word vector dimensionality: Tried various values from 100 to 1000

Context / window size: 10 tokens

Worker threads: 4

Minimum word count: Tried different values in the range of 3-40

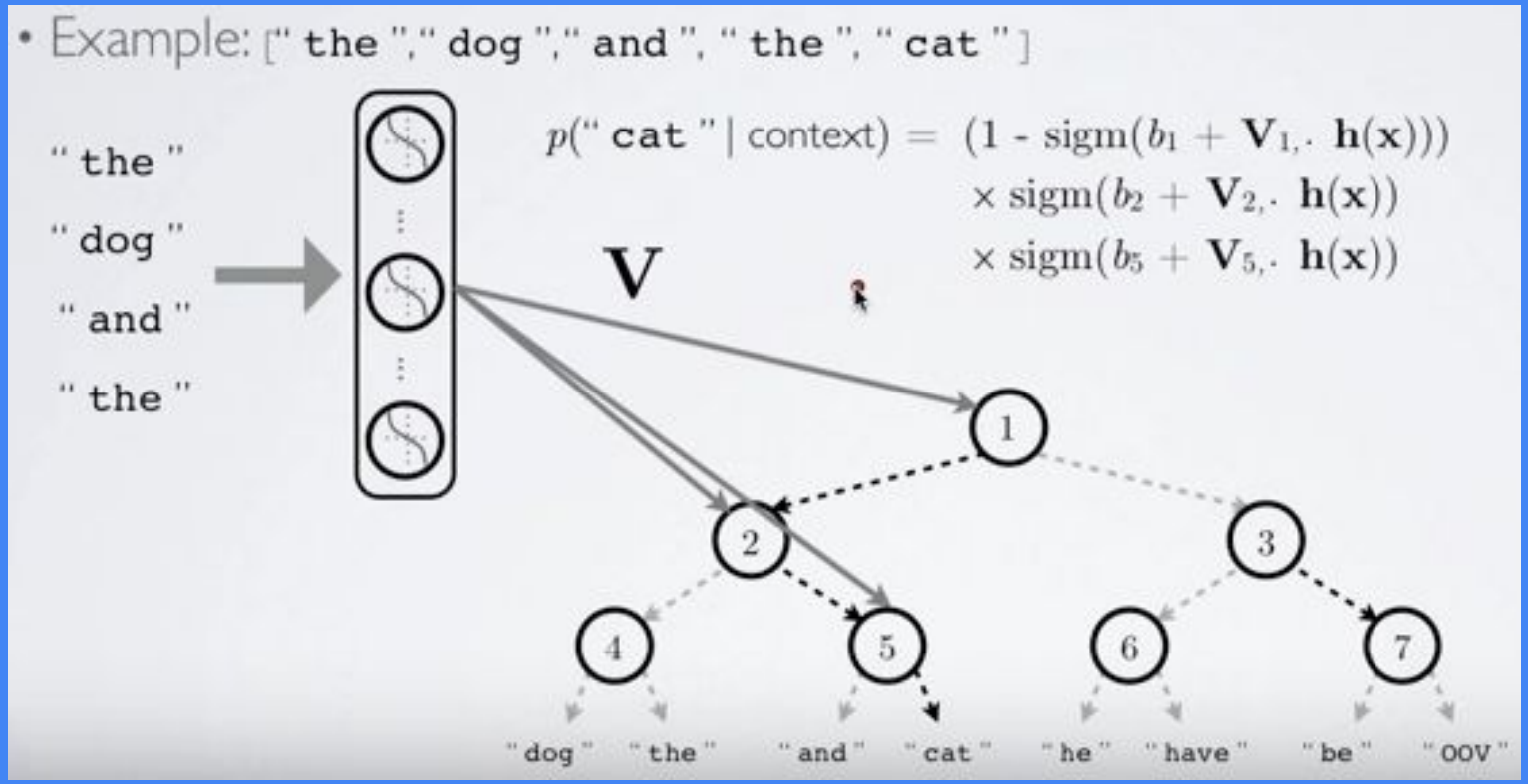
Softmax



$$p(w \mid c) = \frac{\exp(h^\top v'_w)}{\sum_{w_i \in V} \exp(h^\top v'_{w_i})}$$

http://sebastianruder.com/word-embeddings-softmax/

Hierarchical Softmax



http://sebastianruder.com/word-embeddings-softmax/

Outcomes

Unlabeled train set: 2.5 mln lemmatized tokens

Labeled train&test sets: 300 sentence-length contexts

Дані classifier
Accuracy: ~62%

Machine-translated text classifier
Accuracy: ~60%

Word embeddings add valuable functionality

Lessons Learned

- **Need for data** - a good Ukrainian corpus
- Improved data preprocessing
- Word2Vec works, but there are other options to be explored
- Tokens work, but character-based vectors are well worth considering

Vectors Must Go On

Nataliia Cheilytko
natalia.cheilytko@gmail.com LinkedIn

Vasyl Starko
vstarko@gmail.com LinkedIn

Mykhailo Kotov
mykhailo.kotov@gmail.com LinkedIn

Ukrainian Brown Corpus Group

r2u.org.ua/corpus bruk.group@gmail.com facebook.com/r2u.org.ua