AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Narodowe Centrum
Badań i Rozwoju

Contextual Word Embeddings

Krzysztof Wróbel,
Aleksander Smywiński-Pohl

Lemkin - intelligent legal information system

dr inż. Aleksander Smywiński-Pohl

adw. dr Marek Strzała

mgr inż. Krzysztof Wróbel

mgr inż. Mateusz Piech

mgr inż. Klaudia Bałazy

Problem statement

USTAWA

z dnia 6 grudnia 2018 r.

o zmianie ustawy o przekształceniu prawa użytkowania wieczystego gruntów zabudowanych na cele mieszkaniowe w prawo własności tych gruntów

Art. 1. W ustawie z dnia 20 lipca 2018 r. o przekształceniu prawa użytkowania wieczystego gruntów zabudowanych na cele mieszkaniowe w prawo własności tych gruntów (Dz. U. poz. 1716) wprowadza się następujące zmiany:

1) w art. 4:

a) ust. 2 otrzymuje brzmienie:

„2. Organ, o którym mowa w ust. 1, zwany dalej „właściwym organem", wydaje zaświadczenie:

1) z urzędu – nie później niż w terminie 12 miesięcy od dnia przekształcenia albo

2) na wniosek właściciela – w terminie 4 miesięcy od dnia otrzymania wniosku, albo

3) na wniosek właściciela lokalu uzasadniony potrzebą dokonania czynności prawnej mającej za przedmiot lokal albo właściciela gruntu uzasadniony potrzebą ustanowienia odrębnej własności lokalu – w terminie 30 dni od dnia otrzymania wniosku.",

b) po ust. 2 dodaje się ust. 2a w brzmieniu:

„2a. W przypadku, o którym mowa w art. 2 ust. 2, właściwy organ wydaje zaświadczenie w terminie 4 miesięcy od dnia przedstawienia przez cudzoziemca w rozumieniu art. 1 ust. 2 ustawy z dnia 24 marca 1920 r. o nabywaniu nieruchomości przez cudzoziemców, zwanego dalej „cudzoziemcem", ostatecznego zezwolenia, o którym mowa w tym przepisie.";

Corpora features

- vocabulary size
  - NKJP: 143 thousand
  - Full NKJP: 9,4 mln
  - Legal corpus: 784 thousand
- corpus size
  - NKJP: 1,2 mln tokens
  - Full NKJP: 2,2 bln tokens
  - Legal corpus: 4,1 bln tokens

1-hot vs. word embeddings

- 1-hot encoding - the vector size equals the vocabulary size
- word embeddings = dense vector representations of words, vector size is much smaller (e.g. 300)
- unsupervised learning
  - easily adaptable to new languages and problems
- popular systems: word2vec, GloVe, fastText produce so-called static word embeddings
  - representation is independent from the word context and unable to capture ambiguity

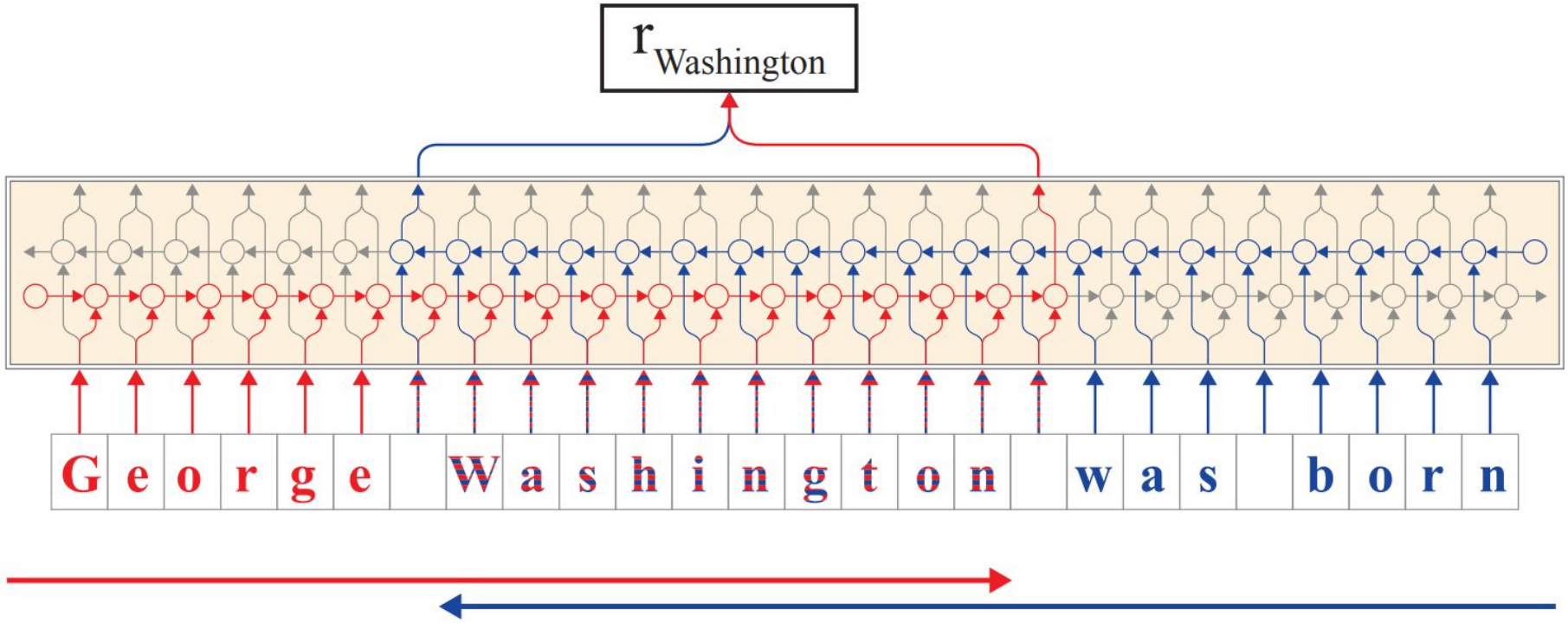
Contextual word embeddings

- solves ambiguity/polysemy problem
- the most recent systems: ELMo, BERT and Flair encode not only the word in question but also its surroundings.
- SOTA architecture in question answering, machine translation, language modeling, text summarization and more

Flair

- processes raw text - does not employ tokenization
- character based embeddings
  - dense representation for an unrestricted span of text
- character language model using recurrent neural networks (LSTM)
  - forward and backward model

Flair



Source: Contextual String Embeddings for Sequence Labeling. Alan Akbik, Duncan Blythe and Roland Vollgraf. 27th International Conference on Computational Linguistics, COLING 2018.

Language model training

- prepare corpus
- define character dictionary
- set parameters
  - learning rate starts from 20.0
  - size of internal state of RNN (LSTM): 1024 or 2048
- code modified to use multiple GPUs
  - 33% speedup using two GPUs on Prometheus
- model size 70MB

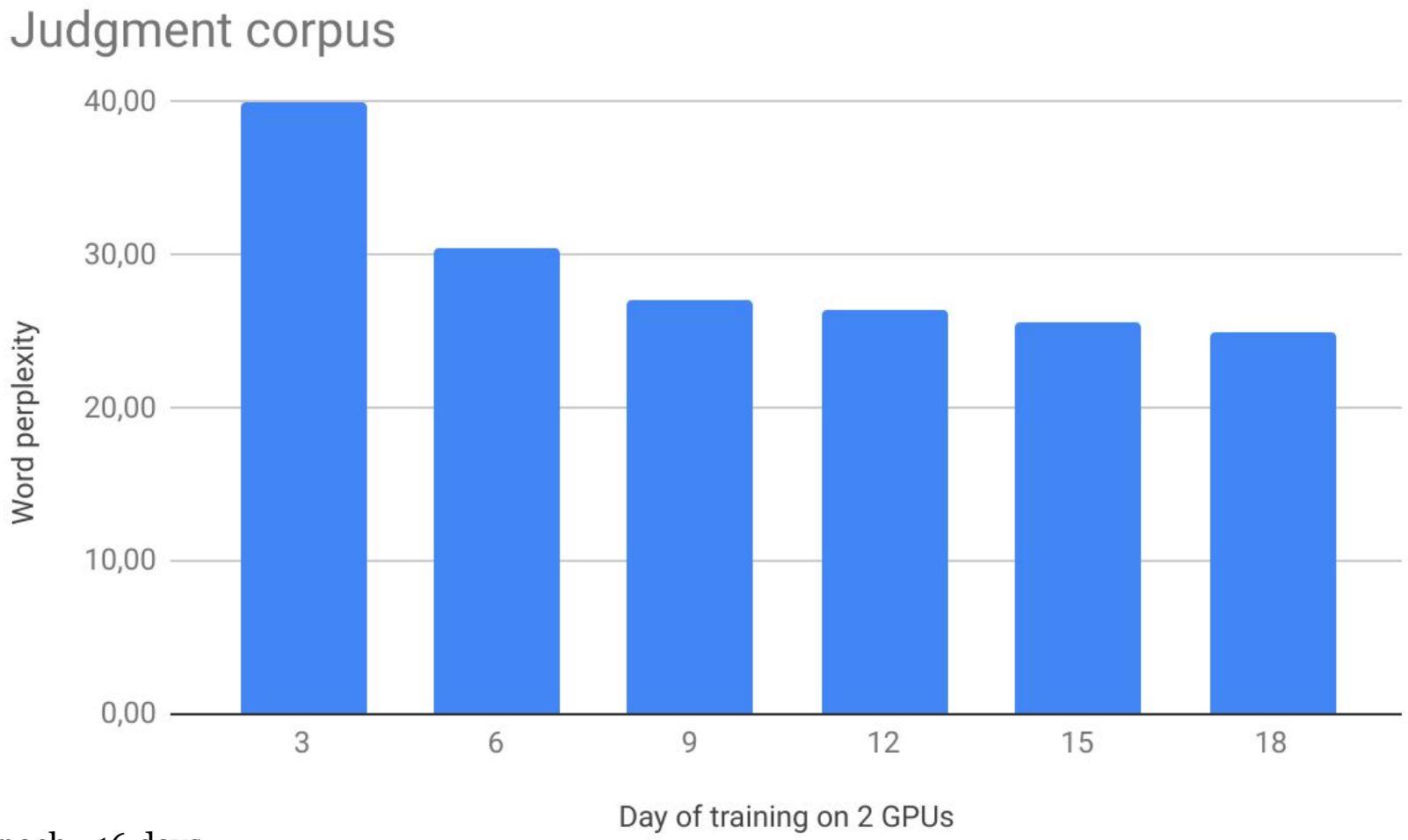
LM - Results

The language models were assessed on separate judgments testing data.

	Character perplexity	**Word perplexity**
Common Crawl	3.42	5527.94
Judgments corpus	1.7	41.07

Word perplexity of 41 - the model is as confused on test data as if it had to choose uniformly and independently among 41 possibilities for each word.

Perplexity reduction



1 epoch - 16 days

Text generation with judgment model

Zakopane członkowie Spółki w taki sposób aby skarżący spłacała jego powyższy zadłużenie. Nie da się zresztą kierować się wskazaniami wyroku Sądu Okręgowego z dnia 18 września 2008 r., w którym Sąd Okręgowy nakazał wpisanie z wiadomością do unijnego Instrumenta Telekomunikacyjnego w P. nie wyczynały się od ostatniego współwłaściciela rzeczy na spornych udziałach Spółki (...) oraz nastąpiło przelewy wierzytelności przejętych przez spółkę P. pod nazwą: (...) T S.A. dla dalej Osobę „Paka'i". Poglądu tego nie uprawdopodobniono w uzasadnieniu zaskarżonego wyroku, niewarunkowania likwidacją wadliwością art. 423 k.c. Przyjęcie poglądu prezentowanego w doktrynie prawa i orzecznictwa jest uzasadnione tylko

Long dependencies

Ona jest bardzo pięknie otoczon**a** twoim wichrzem.

On jest bardzo piękny w taki sposób, ale to nie on.

Sentence Boundary Detection

Goal: segment a text into sentences

- question marks and exclamation marks are omitted (not present in judgments)
- classifier - one layer perceptron
- input: embedding of a full stop generated by Flair

SBD - Results

For sentence boundary detection we performed 5-fold cross-validation on two manually annotated data.

F1 scores:

	Judgment corpus	NKJP
Toki (rule based)	92.2 (manually updated rules: 99.2)	96.2
Our neural network	99.4	99.8

Our model obtained better results than Toki up to 7 percentage points, but Toki does not utilize training data. Also, the neural network is 20 times slower than rule-based Toki.

We plan to parallelize calculations to more nodes and employ Flair embeddings to further tasks in NLP pipeline: tokenization, part-of-speech tagging, detection of more sophisticated phenomena.

AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Narodowe Centrum
Badań i Rozwoju

Contextual Word Embeddings

Krzysztof Wróbel
krzysztof@wrobel.pro,
Aleksander Smywiński-Pohl
apohllo@agh.edu.pl