Native Language Identification

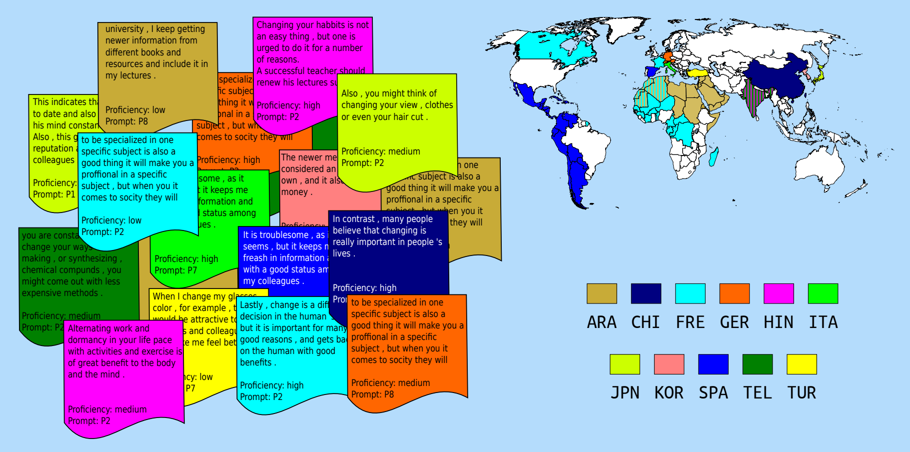
A challenging task description

Barbora Hladká **Martin Holub**

{hladka | holub}@ufal.mff.cuni.cz

Charles University,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Native Language Identification



Native Language Identification

Sample texts

(1) Many people lose thier life , thier job , or maybe thier family . All of them were not learn from other 's fault . Successful person who has many choses that help him in his life . So , what we lose if we learn one more thing in our life .

(2) Last week I read an article in our daily newspaper about who enjoys life more , female or male . I considered about these piece of paper a very long time and came to the conclusion that it is more worth to distinguish between young and old people , than between female and male .

Native Language Identification

Sample texts

(1) Many people lose thier life , thier job , or maybe thier family . All of them were not learn from other 's fault . Successful person who has many choses that help him in his life . So , what we lose if we learn one more thing in our life .

(2) Last week I read an article in our daily newspaper about who enjoys life more , female or male . I considered about these piece of paper a very long time and came to the conclusion that it is more worth to distinguish between young and old people , than between female and male .

(1) . . . **ARA**
(2) . . . **GER**

Native Language Identification

Task: Predict L1 of English essays's authors

Data:

- TOEFL11 is a corpus of non-native English writing
  - consists of essays on eight different topics (prompts P1 to P8)
  - written by non-native speakers of three proficiency levels (labels low/medium/high)
  - the essays' authors speak 11 different native languages that should be predicted
  - contains 1,100 tokenized essays per language with an average of 348 word tokens per essay
  - more info can be found in (Blanchard et al., 2013)
- Additionally, all texts have been preprocessed by the Standford POS tagger (Toutanova et al., 2003).

Existing approaches to NLI
State of the art results

System	# of feat.	Acc.*	Approach
Gebre et al., 2013	73,626	84.6	`tf-idf` of unigrams and bigrams of words
Jarvis, Bestgen, Pepper, 2013	400K	84.5	$\{1,2,3\}$-grams of words, lemmas, POS tags, df ≥ 2
Kríž, Holub, Pecina, 2015	55**	82.4	language models using tokens, characters, POS, suffixes

*Acc. – cross-validation results on `train+dtest`

**traditional n-grams are hidden in the language models

Analyzing most discriminative n-grams

A sample of extracted word *n*-grams

n-gram	fr	df	G max		G max_2		G max_3	
,	114,358	8,701	-2,857.8	TEL-ZHO	2,328.7	JPN-TEL	2,239.8	KOR-TEL
i	6,011	1,939	992	ARA-DEU	957.4	ARA-JPN	735.7	ARA
Japan	356	210	982.1	JPN	380.8	JPN-TUR	379.2	JPN-SPA
think	9,883	4,861	-915.1	HIN-ITA	797.6	ITA-TEL	-767.9	HIN-JPN
think that	3,602	2,322	807.7	ITA-TEL	748.5	ITA	506.2	ITA-ZHO
I think that	1,963	1,375	796.3	ITA	523.6	ITA-ZHO	391.8	ITA-KOR
tour	3,810	694	-725.3	ITA-JPN	-644.7	ITA-KOR	-625.3	DEU-JPN
Indeed	282	222	694.2	FRA	288.5	FRA-HIN	282.4	FRA-TEL
in twenty	1,214	703	73.1	FRA-HIN	-62.8	HIN	NA	HIN-JPN

fr ... absolute n-gram frequency in the corpus
df ... document frequency (number of examples containing given n-gram)

G stands for *G*-test statistic.

Our best model (Kríž, Holub, and Pecina, 2015)

Language modeling approach to feature extraction

- We built 11 special language models of English (M_i), each based on the texts with the same L1 language available in the training data.
- Then we compare M_i to a general language model of English (M_G).
- The cross-entropy of text t with empirical n-gram distribution p given a language model M with distribution q is
$$\mathrm{H}(t, M) = -\sum_x p(x) \log q(x).$$
- **Normalized cross-entropy scores – used as features**

$$D_G(t, M_i) = \mathrm{H}(t, M_i) - \mathrm{H}(t, M_G) = -\sum_x p(x) \log \frac{q_i(x)}{q_G(x)},$$

where M_i are the special language models with distributions q_i, and M_G is the general language model with the distribution q_G.

Our best model – error analysis

Aggregated confusion matrix

Sum of 10 confusion matrices obtained in 10-fold cross validation process

	ARA	DEU	FRA	HIN	ITA	JPN	KOR	SPA	TEL	TUR	ZHO
ARA	738	8	30	24	5	16	9	30	10	32	15
DEU	7	836	13	4	20	4	6	20	1	16	1
FRA	31	6	767	0	35	6	3	39	1	15	1
HIN	19	1	1	684	0	2	3	3	194	9	4
ITA	6	8	20	1	761	0	1	58	0	3	2
JPN	7	2	3	3	2	709	109	6	1	8	22
KOR	14	1	3	3	1	120	676	6	2	12	50
SPA	30	21	45	6	69	3	4	720	1	17	7
TEL	4	2	0	157	0	1	0	2	685	2	1
TUR	32	15	11	16	6	11	23	11	5	778	16
ZHO	12	0	7	2	1	28	66	5	0	8	781

Initial experiments with NN

Model	Reference
Paragraph vector model – Distributed Bag-Of-Words	(Le & Mikolov, 2014)
Paragraph vector model – Distributed Memory	(Le & Mikolov, 2014)
Word-to-Vec Inversion	(Taddy, 2015)
Recurrent Neural Network	(Cho et al., 2014)
Convolutional Neural Network	(Kim, 2014)

Goals

Build and tune a deep neural model to beat the state of the art using

- recurrent neural networks
- convolutional neural networks
- ...

Then the NN model(s) should be compared to the traditional SVM classifiers:

- **what is the difference?**
- **is their performance complementary?**

References I

- Blanchard, Daniel and Tetreault, Joel and Higgins, Derrick and Cahill, Aoife and Chodorow, Martin. TOEFL11: *A Corpus of Non-Native English*. ETS Research Report. 2013.
- Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In Proceedings of SSST-8, EighthWorkshop on Syntax, Semantics and Structure in Statistical Translation, pages 103–111, Doha, Qatar, October. Association for Computational Linguistics.
- Gebre, Binyam Gebrekidan et al. Improving Native Language Identification with TF-IDF Weighting. In: *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 216–223, Atlanta, Georgia, 2013.
- Jarvis, Scott and Bestgen, Yves and Peppre, Steve. Maximizing Classification Accuracy in Native Language Identification. In: *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 111-118, Atlanta, Georgia, 2013.
- Kim, Yoon. 2014. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on EMNLP, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.
- Kríž, Vincent and Holub, Martin and Pecina, Pavel. Feature Extraction for Native Language Identification Using Language Modeling. In: *Proceedings of Recent Advances in Natural Language Processing*. Hisarja, Bulgaria, pp. 298-306, 2015.

References II

- Le, Quoc and Mikolov, Tomas. Distributed Representations of Sentences and Documents. In: *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 2014.
- Taddy, Matt. 2015. Document Classification by Inversion of Distributed Language Representations. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Short Papers), pp. 45–49, Beijing, China, 2015.
- Toutanova Kristina, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.