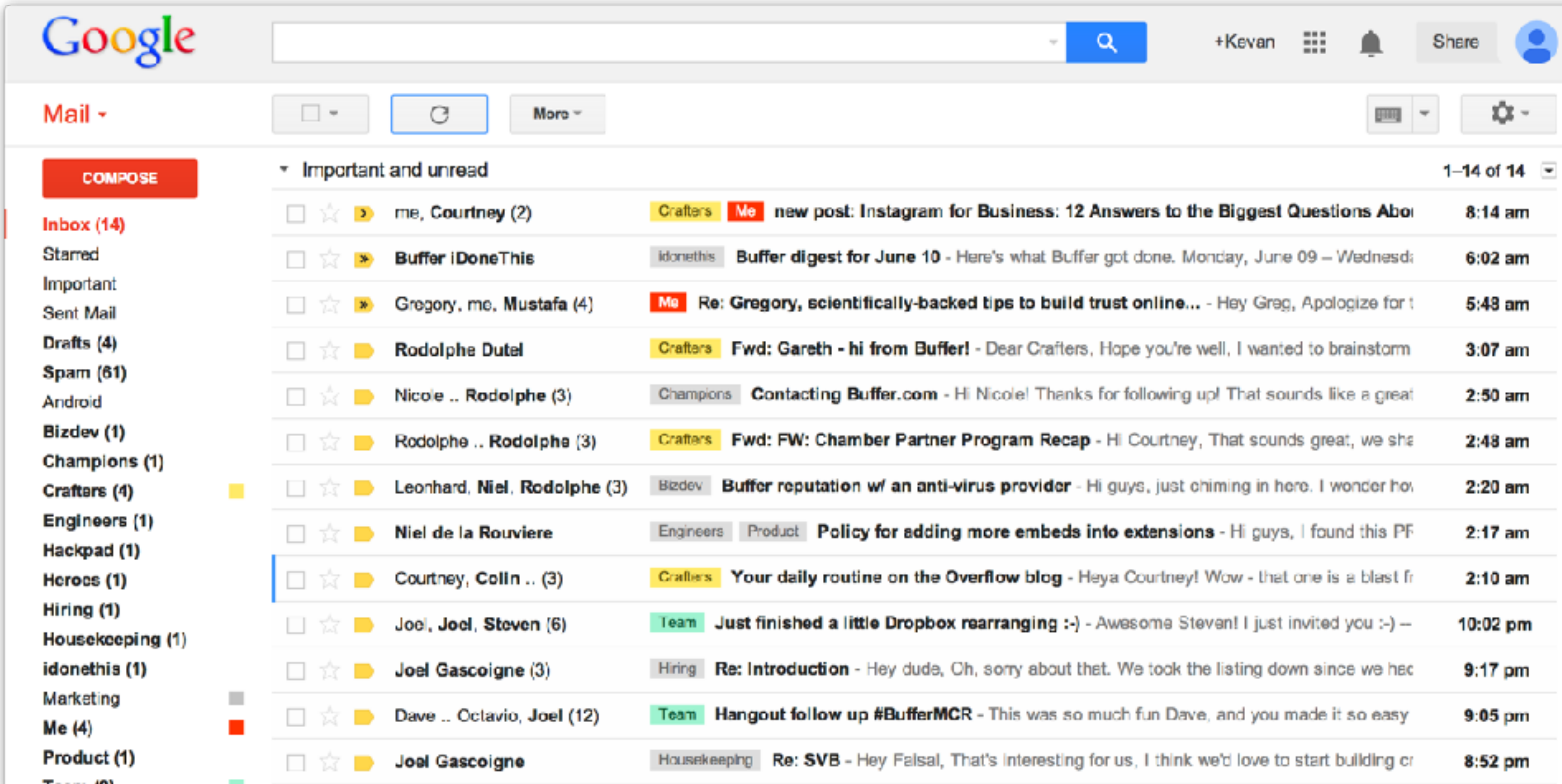
Labeling Text in Several Languages with Multilingual Hierarchical Attention Networks

Nikolaos Pappas, Andrei Popescu-Belis
Idiap Research Institute

June 9, 2017
Swisstext, Winterthur

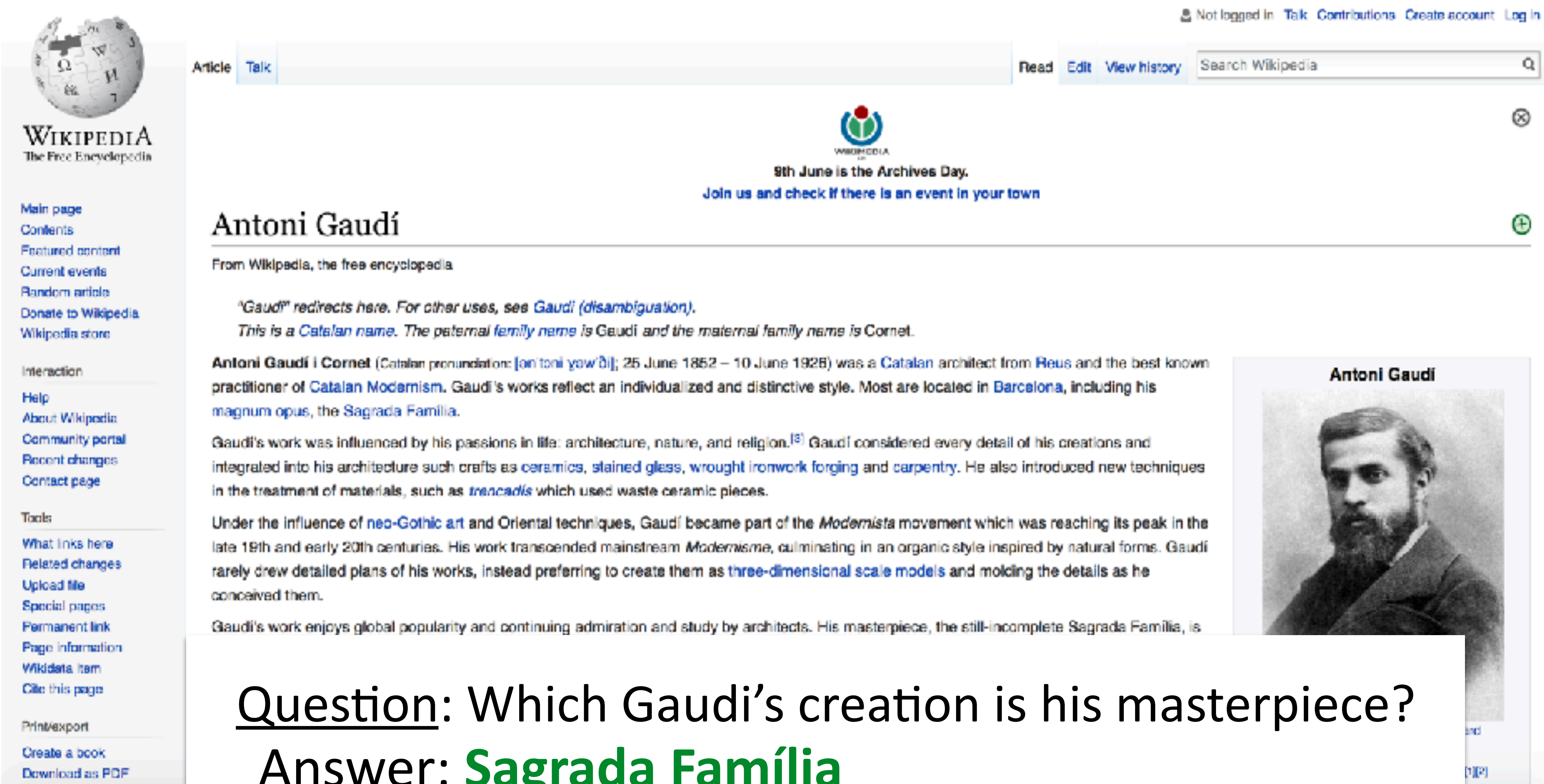
Topic Recognition

Spam filtering — Mailbox Optimization — Customer Support



Question Answering

Reading/Navigation Assistant — Interactive Search

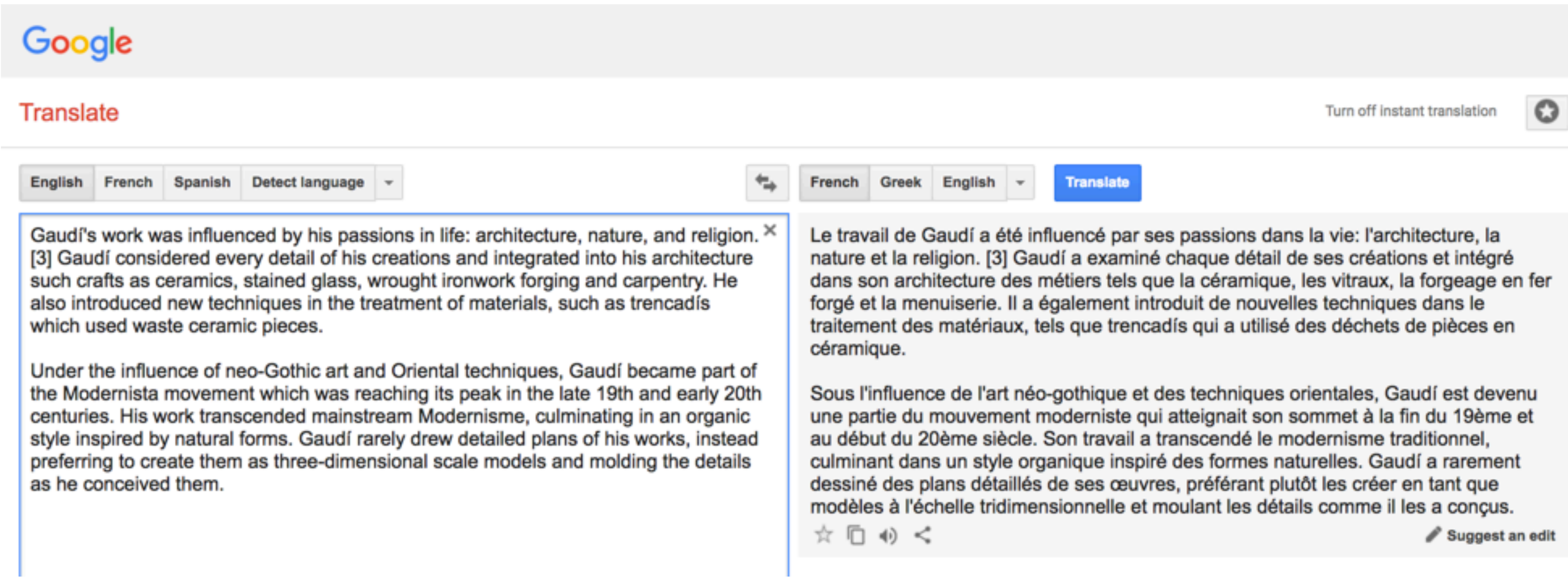


<u>Question</u>: Which Gaudi's creation is his masterpiece?

<u>Answer</u>: **Sagrada Família**

Machine Translation

Document Translation — Dialogue Translation



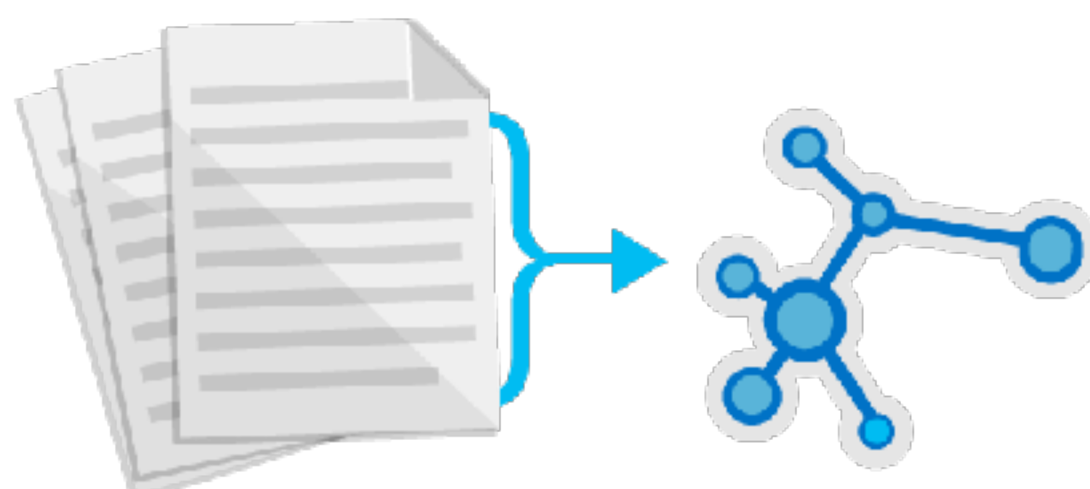
Fundamental Function: Representing Word Sequences

- **Goal:** Learn representations (distributed vectors) of word sequences which encode effectively the meaning / knowledge needed to perform

✓ Topic Recognition
- Question Answering
- Machine Translation
- Summarization

…

$$D = \{(x_i, y_i), i = 1, \dots, N\} \quad y_i \in \{0, 1\}^k$$



Can we benefit from multiple languages?

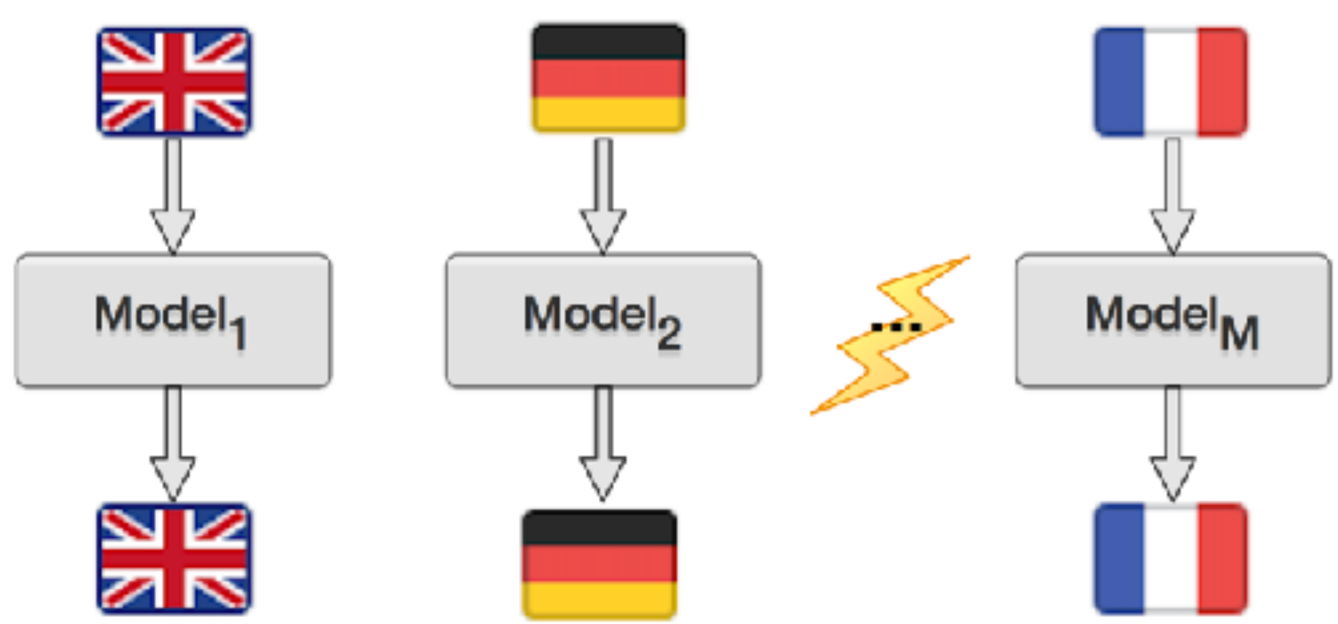
Dealing with Multiple Languages: Monolingually

- **Solution?** Separate models per language
  - language-dependent learning ✗
  - linear growth of the parameters ✗
  - lack of cross-language knowledge transfer ✗
  - hierarchical modeling at the document-level ✓

Documents
$X = \{x_i \mid i=1\ldots n\}$

Models
$f: X \rightarrow Y$

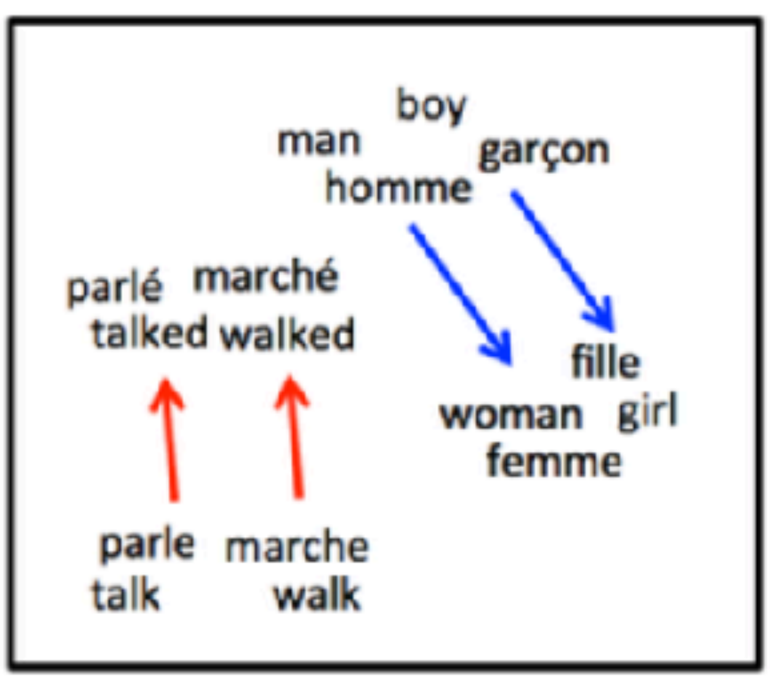
Labels
$Y = \{y_i \mid i=1\ldots n\}$

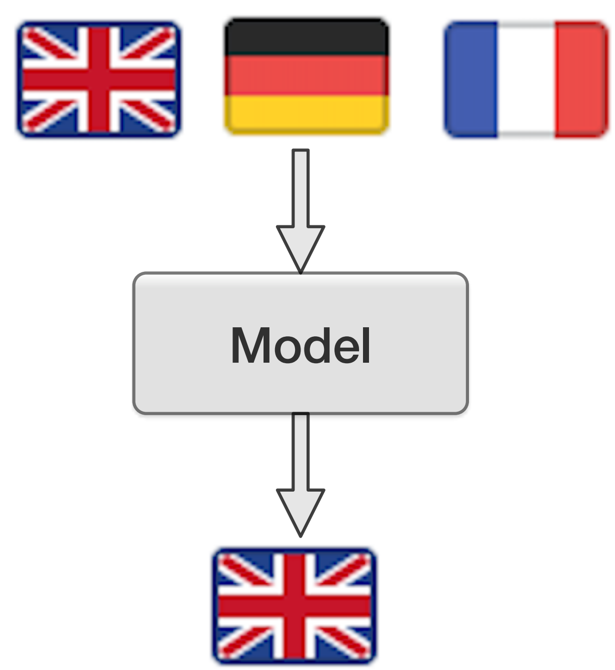


(Kim, 2014)
(Tang et al., 2015)
(Lin et al., 2015)
(Yang et al., 2016)

Dealing with Multiple Languages: Multilingually

- **Solution?** Single model with aligned input space
  - language-independent learning ✓
  - constant number of parameters ✓
  - common label sets across languages ✗
  - modeling at the word-level ✗





(Klementiev et al., 2012)

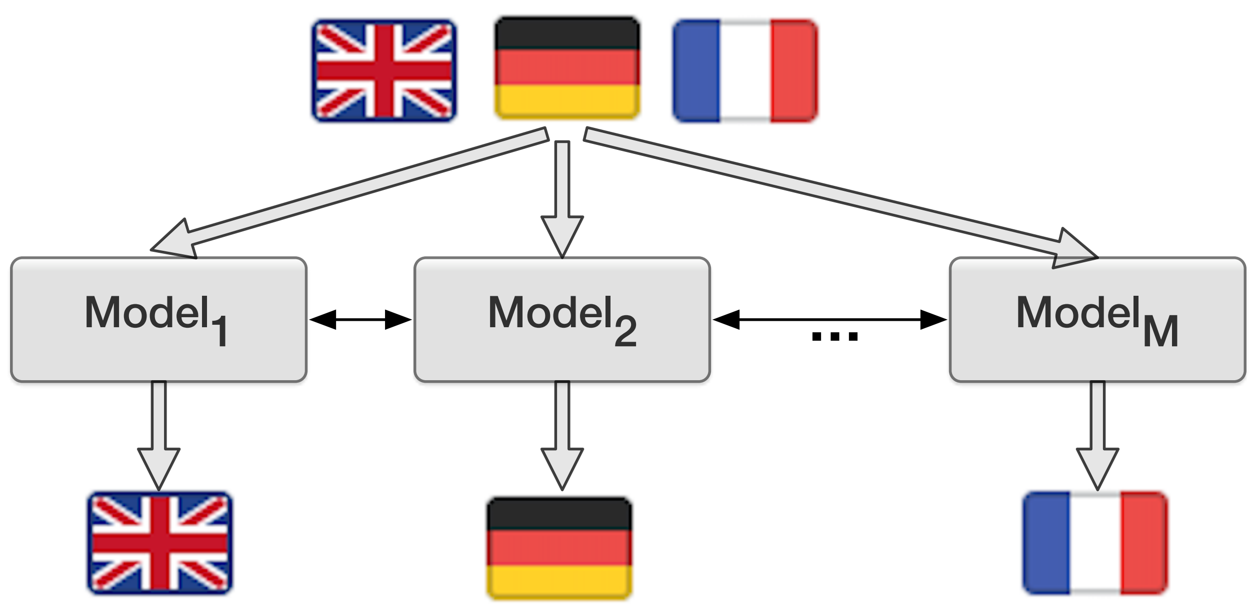
(Herman and Blunsom, 2014)

(Gouws et al., 2015)

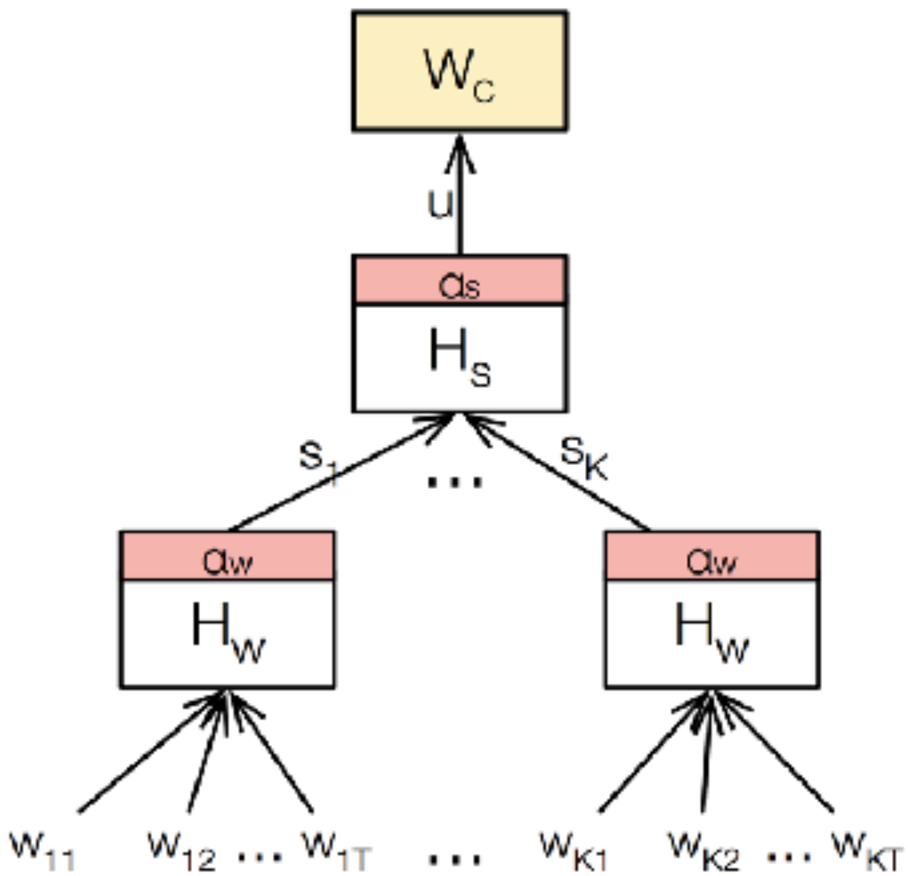
(Ammar et al., 2016)

Dealing with Multiple Languages: Our contribution

- **Solution:** Single model trained over arbitrary label sets with an aligned input space
  - language-independent learning ✓
  - sub-linear growth of parameters ✓
  - arbitrary label sets across languages ✓
  - hierarchical modeling at the document-level ✓



Background: Hierarchical Attention Networks (HANs)



Words: $w_i \in R^d$

Sentences: $s_i \in R^{d_w}$

Document: $u \in R^{d_s}$

- Input: sequence of word vectors

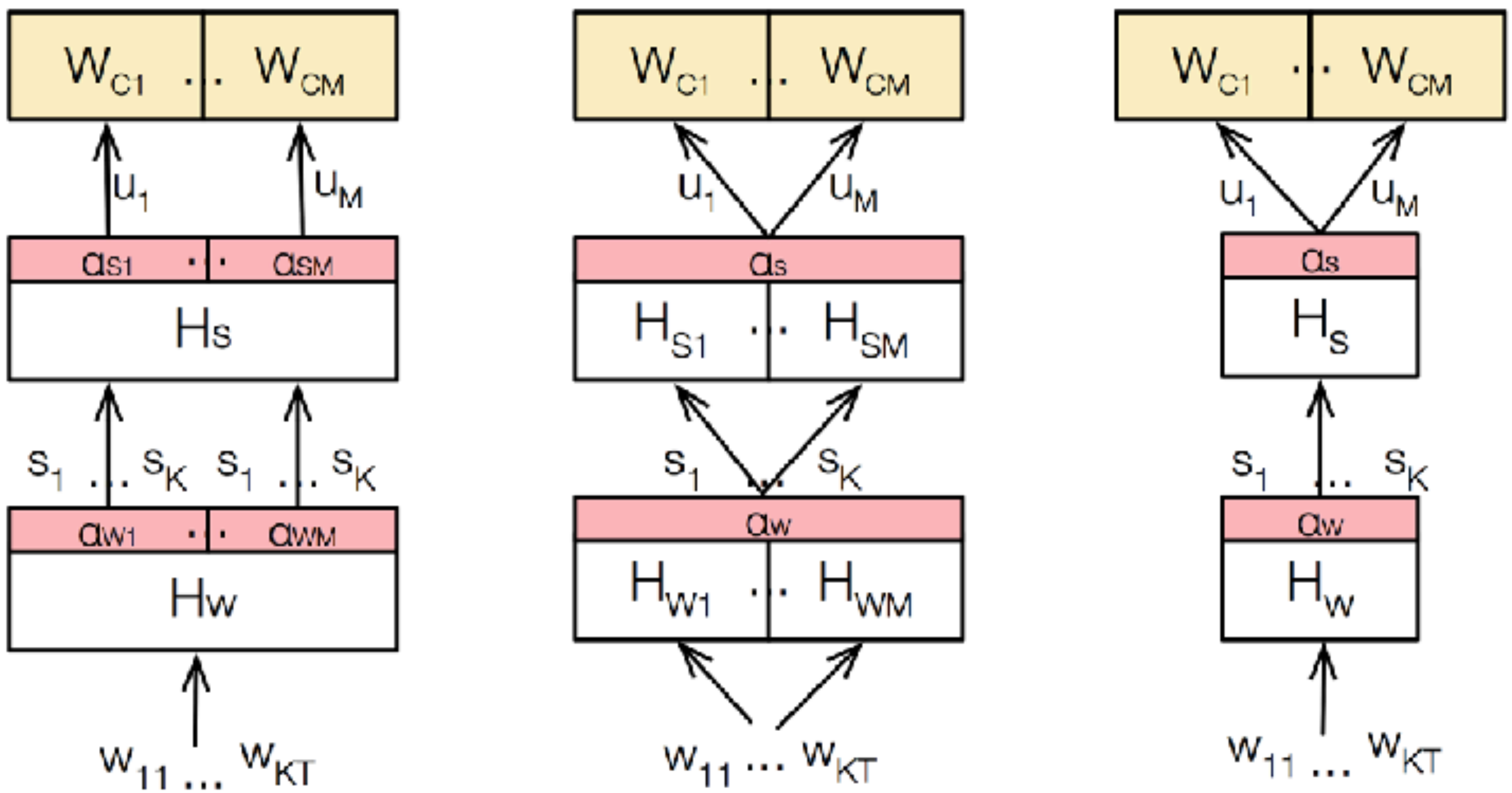
$$x_i = \{w_{11}, w_{12}, \ldots, w_{ST}\}$$

- Output: document vector u

- Hierarchical structure
  - Word-level and sentence-level abstraction layers
    - encoder (Hs, Hw)
    - attention mechanism (aw, αs)
  - Classification layer (Wc) + cross-entropy
- Training: using SGD with ADAM

(Yang et al., 2016)

MHANs: Multilingual Hierarchical Attention Networks



(a) Sharing Encoders (b) Sharing Attentions (c) Sharing Both

Multilingual Attention Networks: Computational Cost

- A fewer number of parameters is needed
  - $\theta_{enc} = \{\mathbf{H}, \mathbf{W}^{(l)}, \mathbf{H}, \mathbf{W}^{(l)}, \mathbf{w}^{(l)}\}$, $\theta_{att} = \{\mathbf{H}^{(l)}, \mathbf{W}, \mathbf{H}^{(l)}, \mathbf{W}, \mathbf{w}^{(l)}\}$
  - $\theta_{both} = \{\mathbf{H}, \mathbf{W}, \mathbf{H}, \mathbf{W}, \mathbf{w}^{(l)}\}$, $\theta_{mono} = \{\mathbf{H}^{(l)}, \mathbf{W}^{(l)}, \mathbf{H}^{(l)}, \mathbf{W}^{(l)}, \mathbf{w}^{(l)}\}$
- The following inequalities are true:

$$|\theta_{mono}| > |\theta_{enc}| > |\theta_{att}| > |\theta_{both}|$$

- Example with shared attention mechanisms

Word emb.	$\|L\|$	$Y_{general}$		$Y_{specific}$	
aligned	1	50K –	77.41 –	90K –	44.90 –
	2	40K ↓	78.30 ↑	80K ↓	45.72 ↑
	8	32K ↓	77.91 ↑	72K↓	45.82 ↑
non-aligned	8	32K ↓	71.23 ↓	72K ↓	33.41 ↓

Naive DL multilingual adaptation fails!

Multilingual Attention Networks: Training Strategy

- Minimizing the sum of the cross-entropy errors

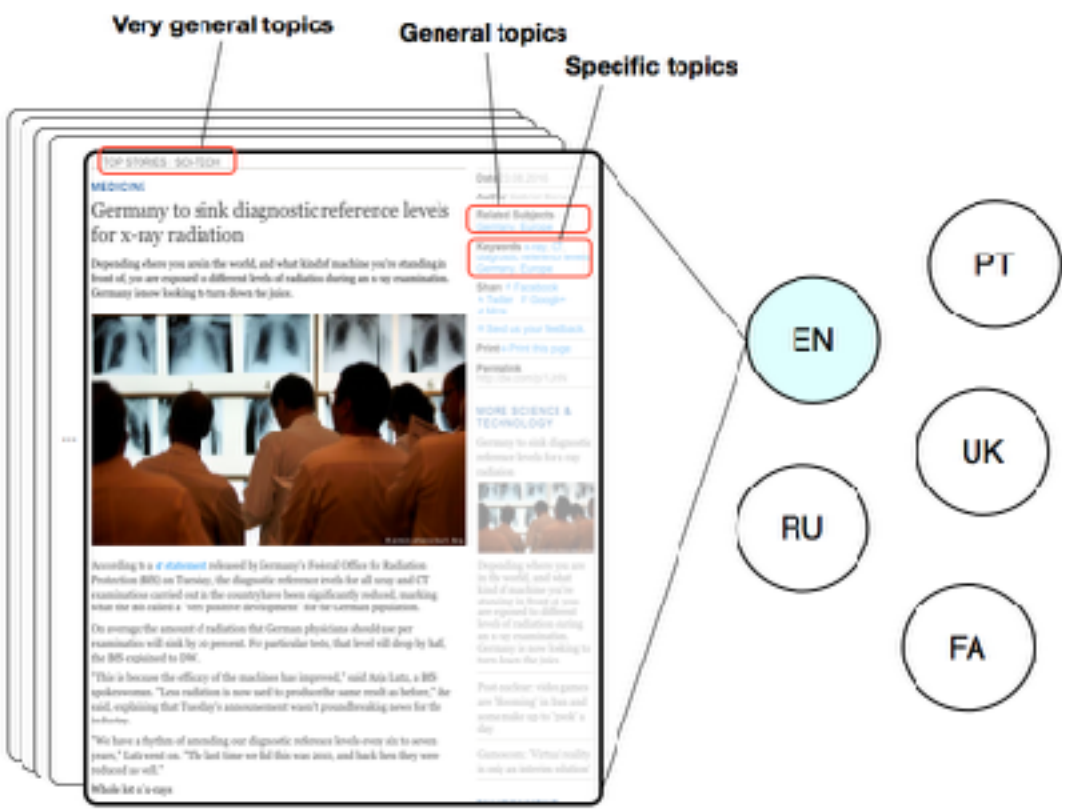
$$\mathcal{L}(\theta_{1,\ldots,}\theta_M) = -\frac{1}{Z}\sum_{l}^{M}\gamma_l\sum_{i}^{N_e}\mathcal{H}(y_i^{(l)}, \hat{y}_i^{(l)}) \quad (8)$$

- **Issue**: Naive consecutive training biases the model
- Sample document-label pairs for each language in a cyclic fashion:

$$(L_1, \ldots, L_M)^{(1)} \rightarrow \ldots \rightarrow (L_1, \ldots, L_M)^{(M)}$$

- **Optimizer**: SGD with ADAM (same as before)

Dataset: Deutsche Welle Corpus (600k docs, 8 langs)



Tagged by **journalists**

Languages	Documents			Labels	
L	$\|X\|$	$\bar{s}$	$\bar{w}$	$\|Y_g\|$	$\|Y_s\|$
English	112,816	17.9	516.2	327	1,058
German	132,709	22.3	424.1	367	809
Spanish	75,827	13.8	412.9	159	684
Portuguese	39,474	20.2	571.9	95	301
Ukrainian	35,423	17.6	342.9	28	260
Russian	108,076	16.4	330.1	102	814
Arabic	57,697	13.3	357.7	91	344
Persian	36,282	18.7	538.4	71	127
All	598,304	17.52	436.7	1,240	4,397

Table 1: Statistics of the Deutsche Welle corpus: $\bar{s}$ and $\bar{w}$ are the average numbers of sentences and words per document.

Full-resource Scenario: Bilingual Training

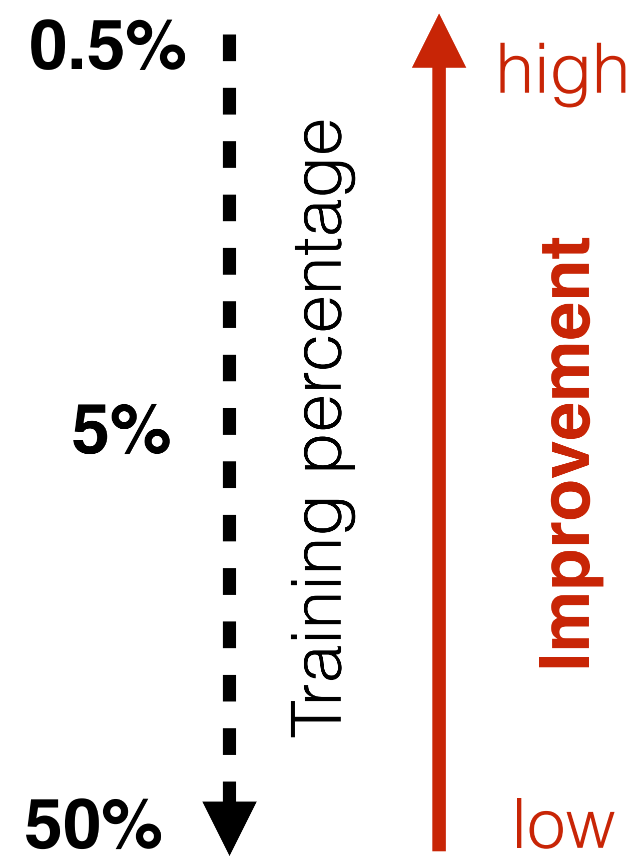
		Models	Auxiliary → English							English → Target						
			de	es	pt	uk	ru	ar	fa	de	es	pt	uk	ru	ar	fa
$Y_{general}$	Mono	NN (Avg)				50.7				53.1	70.0	57.2	80.9	59.3	64.4	66.6
		HNN (Avg)				70.0				67.9	82.5	70.5	86.8	77.4	79.0	76.6
		HAN (Att)				71.2				71.8	82.8	71.3	85.3	79.8	80.5	76.6
	Multi	MHAN-Enc	71.0	69.9	69.2	70.8	71.5	70.0	71.3	69.7	**82.9**	69.7	86.8	80.3	79.0	76.0
		MHAN-Att	**74.0**	**74.2**	**74.1**	**72.9**	**73.9**	**73.8**	**73.3**	**72.5**	82.5	70.8	**87.7**	80.5	**82.1**	76.3
		MHAN-Both	72.8	71.2	70.5	65.6	71.1	68.9	69.2	70.4	82.8	**71.6**	87.5	**80.8**	79.1	**77.1**
$Y_{specific}$	Mono	NN (Avg)				24.4				21.8	22.1	24.3	33.0	26.0	24.1	32.1
		HNN (Avg)				39.3				39.6	37.9	33.6	42.2	39.3	34.6	43.1
		HAN (Att)				43.4				44.8	46.3	41.9	46.4	45.8	41.2	49.4
	Multi	MHAN-Enc	45.4	45.9	44.3	41.1	42.1	44.9	41.0	43.9	46.2	39.3	47.4	45.0	37.9	48.6
		MHAN-Att	**46.3**	**46.0**	**45.9**	**45.6**	**46.4**	**46.4**	**46.1**	**46.5**	**46.7**	**43.3**	**47.9**	45.8	**41.3**	48.0
		MHAN-Both	45.7	45.6	41.5	41.2	45.6	44.6	43.0	45.9	46.4	40.3	46.3	**46.1**	40.7	**50.3**

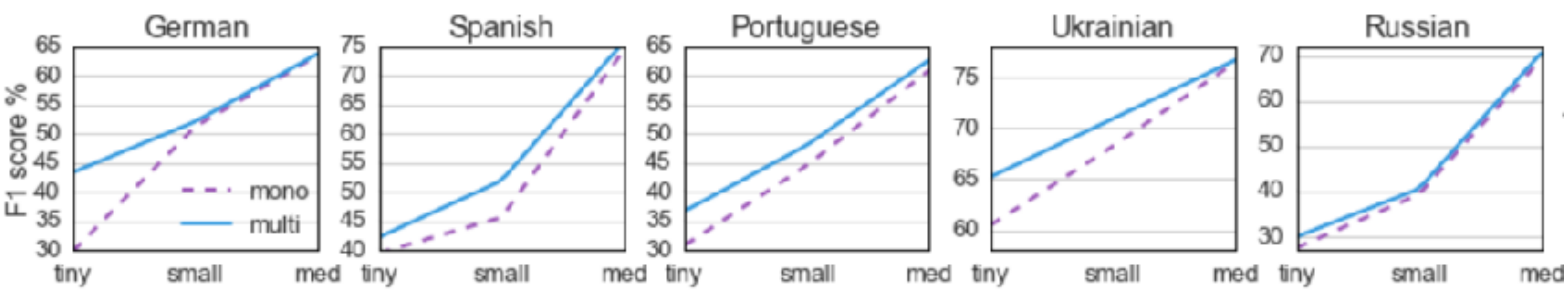
<u>Input</u>: 40-d, <u>Encoders</u>: Dense 100-d, <u>Attentions</u>: Dense 100-d <u>Activation</u>: relu

- Multilingual models consistently outperform monolingual ones
- Sharing attention is the best configuration (on average)
- Traditional (bow) vs neural (en+ar, biGRU encoders)
  - en: 75.8% vs **77.8%** — ar: 81.8% vs **84.0%**

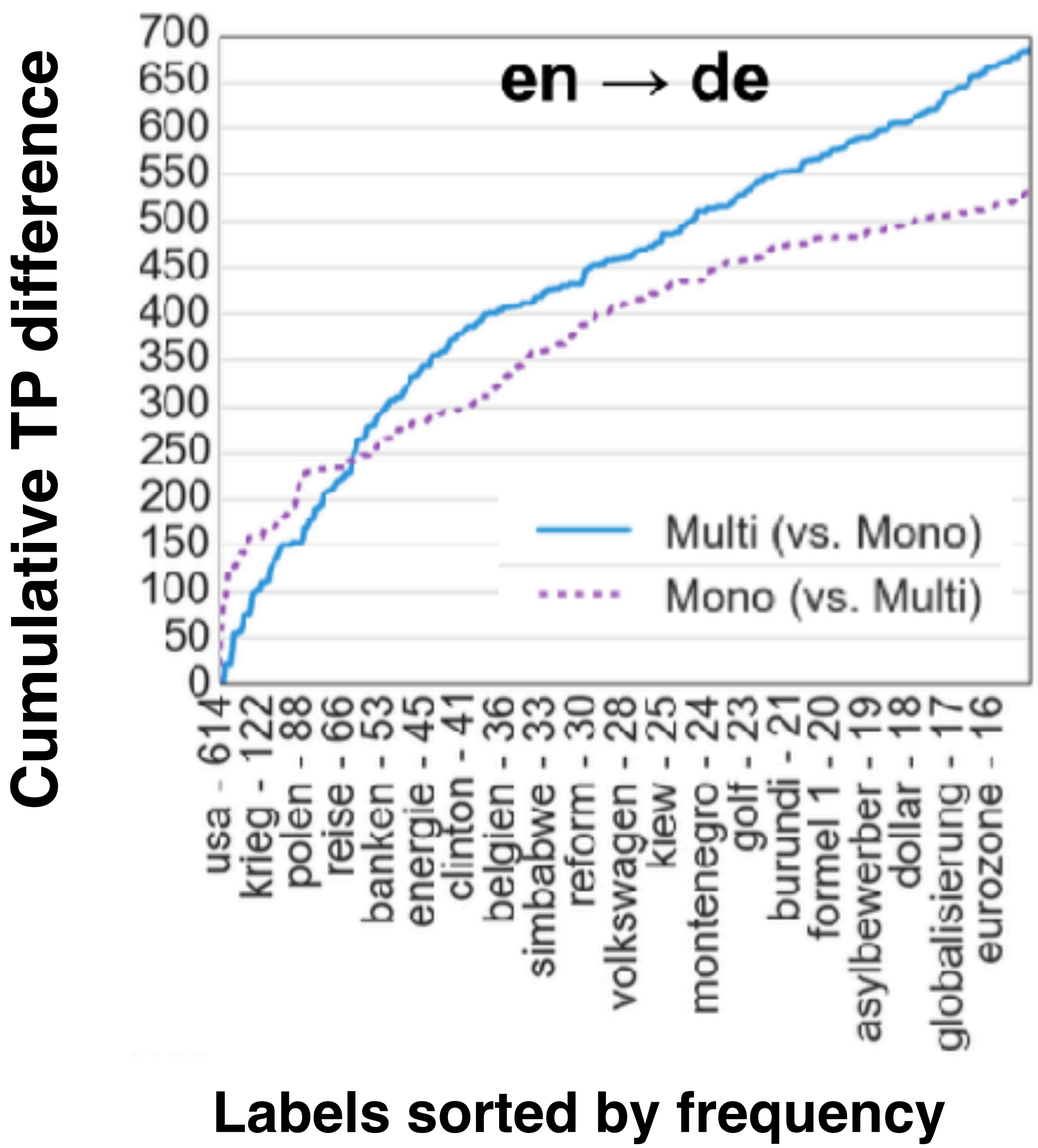
Low-resource Scenario: Bilingual Training

	Size	Mono	Multi			
	$Y_{general}$	HAN	Enc	Att	Both	Δ%
en→de	0.1-0.5%	29.9	**41.0**	37.0	39.4	+37.1
	1-5%	51.3	51.7	49.7	**52.6**	+2.5
	10-50%	63.3	63.0	63.8	**63.8**	+0.7
en→es	0.1-0.5%	38.9	38.6	33.3	**41.5**	+6.7
	1-5%	45.5	50.8	**50.8**	50.1	+11.6
	10-50%	74.2	**75.7**	74.2	75.2	+2.0
en→pt	0.1-0.5%	30.9	25.3	31.6	**33.8**	+9.4
	1-5%	44.6	44.3	37.5	**47.3**	+6.0
	10-50%	60.9	61.9	62.1	**62.1**	+1.9
en→uk	0.1-0.5%	60.4	**62.4**	59.8	60.9	+3.3
	1-5%	68.2	67.7	**70.6**	69.0	+3.5
	10-50%	76.4	76.2	76.3	**76.7**	+0.3





Qualitative Analysis: English - German



- True positive difference (multi vs mono) increases over the entire spectrum
  - German
    russland (21), berlin (19), irak (14), wahlen (13) and nato (13)
  - English
    germany (259), german (97), soccer (73), football 753 (47) and merkel (25)

Qualitative Analysis: Interpretable Output

Conclusion and Perspectives

- New multilingual models to learn shared document structures for text classification
  - Benefit **full-resource** and **low-resource** languages
  - Achieve better accuracy with fewer parameters
  - Capable of cross-language transfer
- Future work
  - Remove the constraint of closed label sets
  - Incorporate label information
  - Apply to other NLU tasks

Thank you

Scalable **U**nderstanding of **M**ultilingual **M**edi**A**

User group meeting
July 3, 2017 Caversham, UK
Demos
Technical talks
Posters & discussions
Contact us if interested!

More about SUMMA:
www.summa-project.eu
info@summa-project.eu

References

- Multilingual Hierarchical Attention Networks for Text Classification, Nikolaos Pappas and Andrei Popescu-Belis, 2017 (submitted)
- Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. CoRR abs/1602.01925.
- Stephan Gouws, Yoshua Bengio, and Gregory S. Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. 32nd International Conference on Machine Learning.
- Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. 52nd Annual Meeting of the Association for Computational Linguistics.
- Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed rep- 894 resentations of words. International Conference on Computational Linguistics.
- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
- Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In Empirical Methods on Natural Language Processing.
- Rui Lin, Shujie Liu, Muyun Yang, Mu Li, Ming Zhou, and Sheng Li. 2015. Hierarchical recurrent neural network for document modeling. Conference on Empirical Methods in Natural Language Processing.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. Conference on Empirical Methods in Natural Language Processing.