



# Labeling Text in Several Languages with Multilingual Hierarchical Attention Networks

**Nikolaos Pappas, Andrei Popescu-Belis**  
Idiap Research Institute



Zürcher Hochschule  
für Angewandte Wissenschaften



June 9, 2017

Swisstext, Winterthur

# Topic Recognition

Spam filtering — Mailbox Optimization — Customer Support

The screenshot shows a Gmail interface with a search bar at the top and a navigation menu on the left. The main area displays a list of 14 important and unread emails. Each email entry includes a checkbox, a star icon, a sender name, a subject line, and a timestamp. The emails are categorized by topics such as 'Crafters', 'Me', 'Team', 'Hiring', and 'Housekeeping'.

Sender	Subject	Time
me, Courtney (2)	new post: Instagram for Business: 12 Answers to the Biggest Questions Abou	8:14 am
Buffer   DoneThis	Buffer digest for June 10 - Here's what Buffer got done. Monday, June 09 – Wednesd	6:02 am
Gregory, me, Mustafa (4)	Re: Gregory, scientifically-backed tips to build trust online... - Hey Greg, Apologize for t	5:48 am
Rodolphe Dutel	Fwd: Gareth - hi from Buffer! - Dear Crafters, Hope you're well, I wanted to brainstorm	3:07 am
Nicole .. Rodolphe (3)	Contacting Buffer.com - Hi Nicole! Thanks for following up! That sounds like a great	2:50 am
Rodolphe .. Rodolphe (3)	Fwd: FW: Chamber Partner Program Recap - Hi Courtney, That sounds great, we sha	2:48 am
Leonhard, Niel, Rodolphe (3)	Buffer reputation w/ an anti-virus provider - Hi guys, just chiming in here. I wonder ho	2:20 am
Niel de la Rouviere	Policy for adding more embeds into extensions - Hi guys, I found this PF	2:17 am
Courtney, Collin .. (3)	Your daily routine on the Overflow blog - Heya Courtney! Wow - that one is a blast fr	2:10 am
Joel, Joel, Steven (6)	Just finished a little Dropbox rearranging :-). Awesome Steven! I just invited you :-). --	10:02 pm
Joel Gascoigne (3)	Re: Introduction - Hey dude, Oh, sorry about that. We took the listing down since we hac	9:17 pm
Dave .. Octavio, Joel (12)	Hangout follow up #BufferMCR - This was so much fun Dave, and you made it so easy	9:05 pm
Joel Gascoigne	Re: SVB - Hey Faisal, That's interesting for us, I think we'd love to start building cr	8:52 pm

# Question Answering

## Reading/Navigation Assistant — Interactive Search

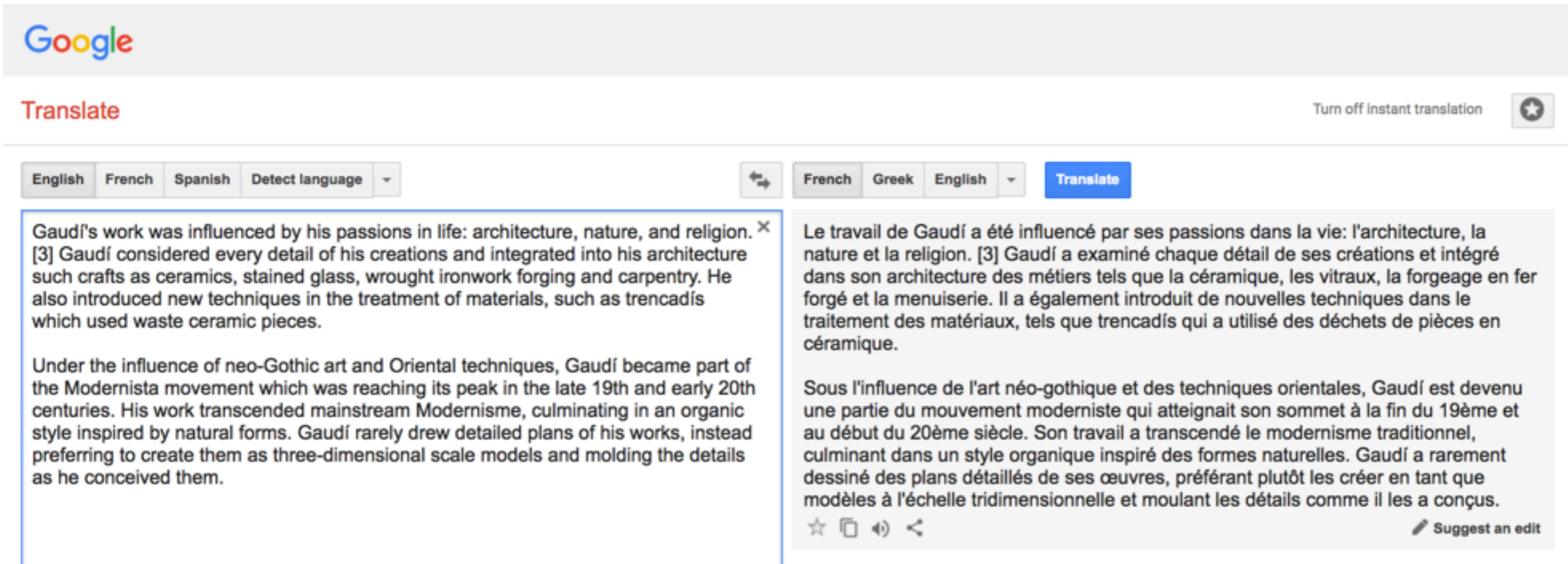


The image shows a screenshot of the Wikipedia article for Antoni Gaudí. The page includes the Wikipedia logo, navigation tabs (Article, Talk), and a search bar. The article text describes Gaudí as a Catalan architect from Reus, known for his work in Catalan Modernism, and mentions his magnum opus, the Sagrada Família. A portrait of Gaudí is visible on the right side of the page. Overlaid on the bottom of the screenshot is a white box containing a question and its answer.

**Question:** Which Gaudí's creation is his masterpiece?  
**Answer:** **Sagrada Família**

# Machine Translation

## Document Translation — Dialogue Translation



The screenshot shows the Google Translate interface. At the top left is the Google logo. Below it, the word "Translate" is written in red. On the right side, there is a link "Turn off instant translation" and a star icon. The main interface has two language selection boxes. The first box on the left contains "English", "French", "Spanish", and "Detect language" with a dropdown arrow. The second box on the right contains "French", "Greek", and "English" with a dropdown arrow. A blue "Translate" button is positioned between the two boxes. Below the language boxes, there are two text areas. The left area contains the English text: "Gaudí's work was influenced by his passions in life: architecture, nature, and religion. [3] Gaudí considered every detail of his creations and integrated into his architecture such crafts as ceramics, stained glass, wrought ironwork forging and carpentry. He also introduced new techniques in the treatment of materials, such as trencadís which used waste ceramic pieces. Under the influence of neo-Gothic art and Oriental techniques, Gaudí became part of the Modernista movement which was reaching its peak in the late 19th and early 20th centuries. His work transcended mainstream Modernisme, culminating in an organic style inspired by natural forms. Gaudí rarely drew detailed plans of his works, instead preferring to create them as three-dimensional scale models and molding the details as he conceived them." The right area contains the French translation: "Le travail de Gaudí a été influencé par ses passions dans la vie: l'architecture, la nature et la religion. [3] Gaudí a examiné chaque détail de ses créations et intégré dans son architecture des métiers tels que la céramique, les vitraux, la forgeage en fer forgé et la menuiserie. Il a également introduit de nouvelles techniques dans le traitement des matériaux, tels que trencadís qui a utilisé des déchets de pièces en céramique. Sous l'influence de l'art néo-gothique et des techniques orientales, Gaudí est devenu une partie du mouvement moderniste qui atteignait son sommet à la fin du 19ème et au début du 20ème siècle. Son travail a transcendé le modernisme traditionnel, culminant dans un style organique inspiré des formes naturelles. Gaudí a rarement dessiné des plans détaillés de ses œuvres, préférant plutôt les créer en tant que modèles à l'échelle tridimensionnelle et moulant les détails comme il les a conçus." Below the French text, there are icons for star, copy, speaker, and left arrow, and a link "Suggest an edit".



# Fundamental Function: Representing Word Sequences

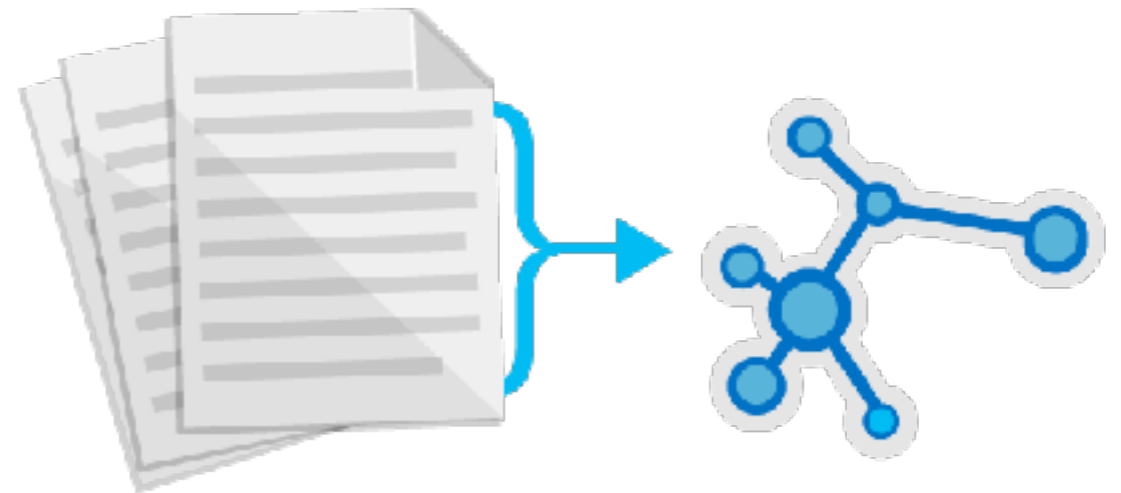
- **Goal:** Learn representations (distributed vectors) of word sequences which encode effectively the meaning / knowledge needed to perform

✓ Topic Recognition

- Question Answering
- Machine Translation
- Summarization

...

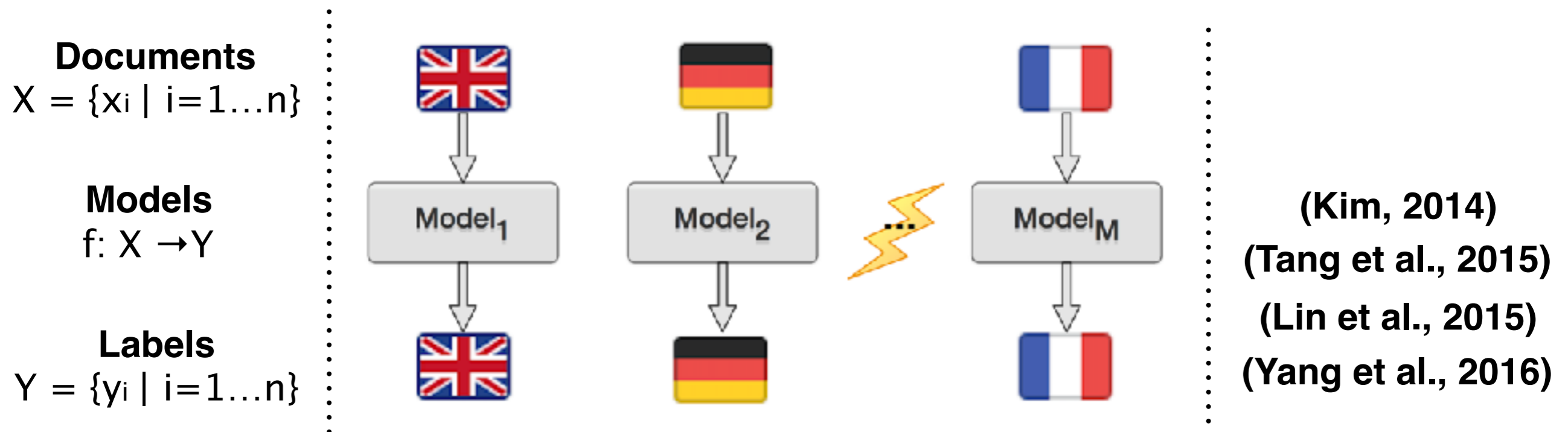
$$D = \{(x_i, y_i), i = 1, \dots, N\} \quad y_i \in \{0, 1\}^k$$



**Can we benefit from multiple languages?**

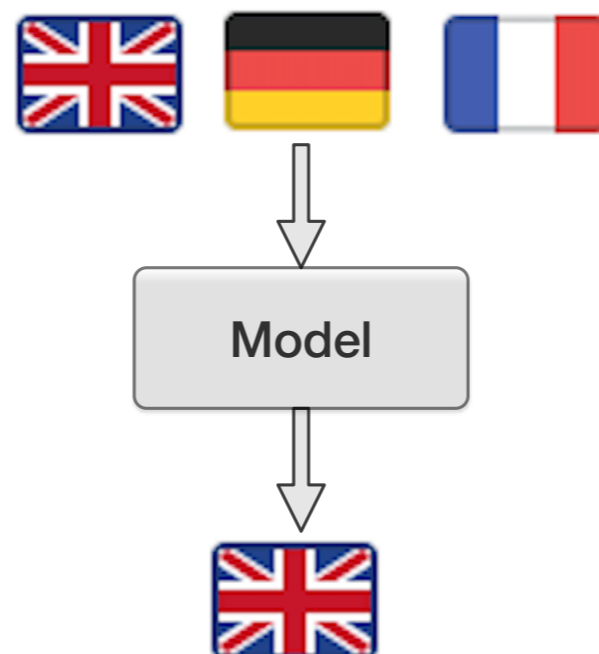
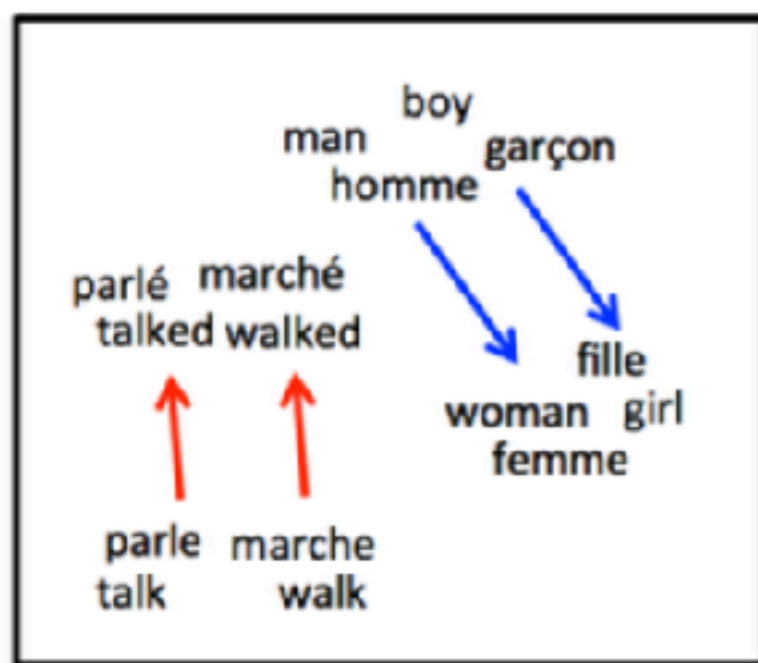
# Dealing with Multiple Languages: Monolingually

- **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 ✅



# 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)

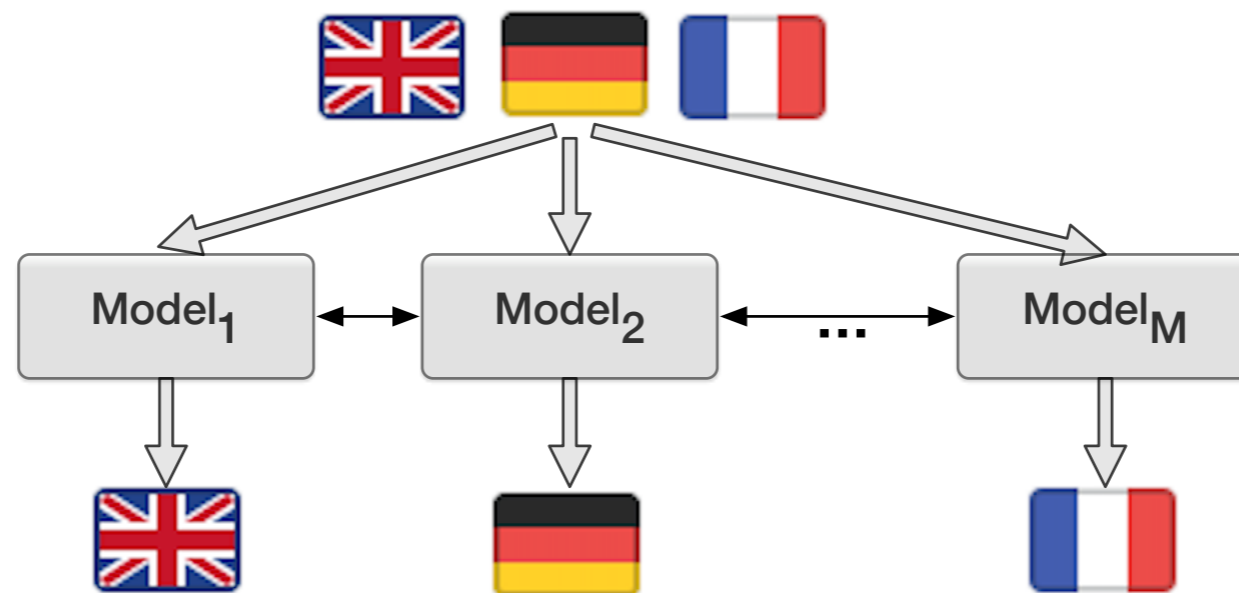
(Herman and Blunsom, 2014)

(Gouws et al., 2015)

(Ammar et al., 2016)

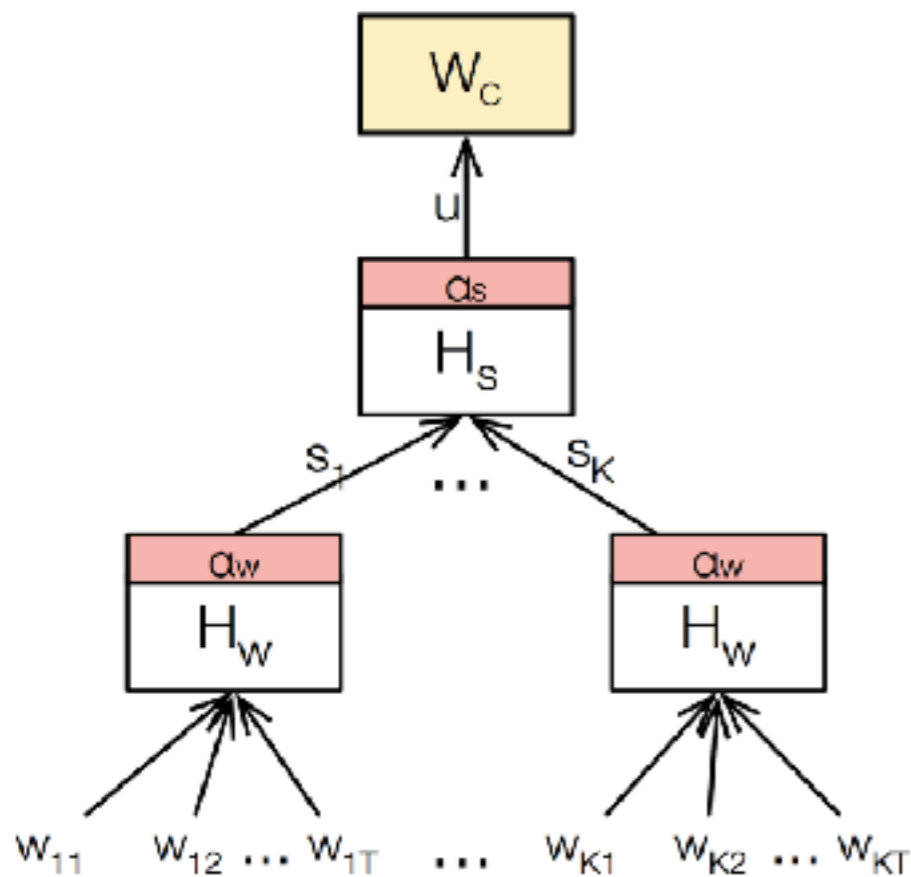
# Dealing with Multiple Languages: Our contribution

- **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:  $w_i \in \mathbb{R}^d$

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

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

- Input: sequence of word vectors

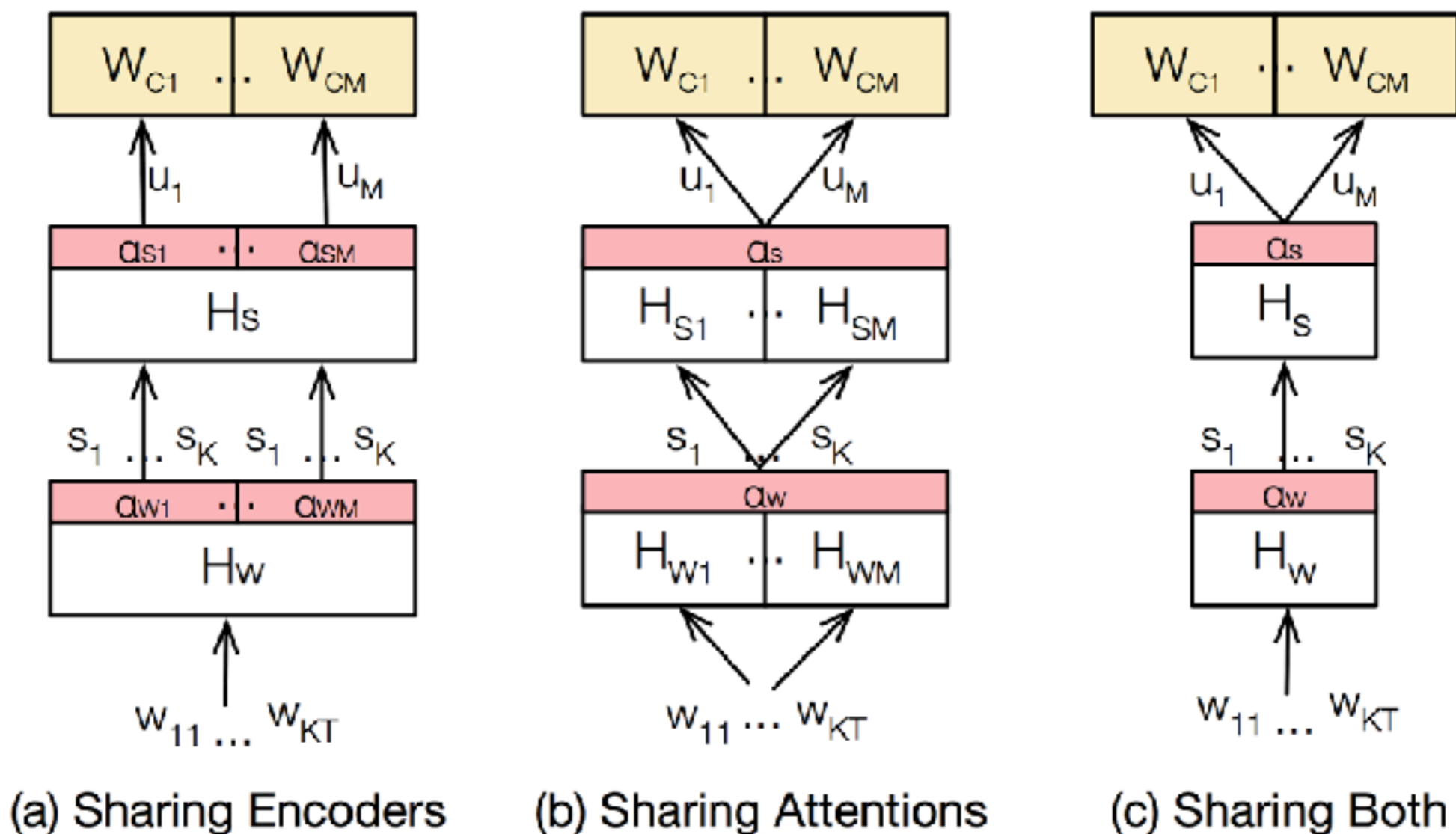
$$x_i = \{w_{11}, w_{12}, \dots, w_{ST}\}$$

- Output: document vector  $u$

- Hierarchical structure
  - Word-level and sentence-level abstraction layers
  - encoder ( $H_s, H_w$ )
  - attention mechanism ( $\alpha_w, \alpha_s$ )
  - Classification layer ( $W_c$ ) + cross-entropy
- Training: using SGD with ADAM

(Yang et al., 2016)

# MHANs: Multilingual Hierarchical Attention Networks



# Multilingual Attention Networks: Computational Cost

- A fewer number of parameters is needed
  - $\theta_{enc} = \{\mathbf{H}, \mathbf{W}^{(1)}, \mathbf{H}, \mathbf{W}^{(1)}, \mathbf{W}^{(1)}\}$ ,  $\theta_{att} = \{\mathbf{H}^{(1)}, \mathbf{W}, \mathbf{H}^{(1)}, \mathbf{W}, \mathbf{W}^{(1)}\}$
  - $\theta_{both} = \{\mathbf{H}, \mathbf{W}, \mathbf{H}, \mathbf{W}, \mathbf{W}^{(1)}\}$ ,  $\theta_{mono} = \{\mathbf{H}^{(1)}, \mathbf{W}^{(1)}, \mathbf{H}^{(1)}, \mathbf{W}^{(1)}, \mathbf{W}^{(1)}\}$
- 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

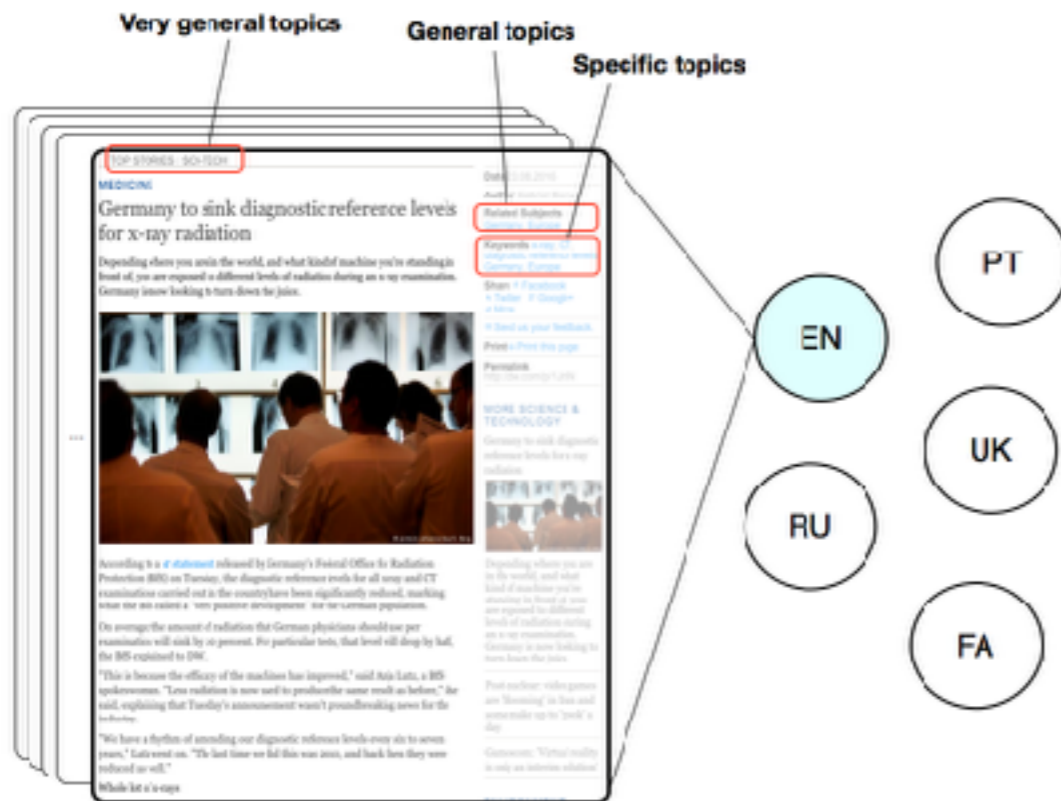
$$\mathcal{L}(\theta_1, \dots, \theta_M) = -\frac{1}{Z} \sum_l^M \gamma_l \sum_i^{N_e} \mathcal{H}(y_i^{(l)}, \hat{y}_i^{(l)}) \quad (8)$$

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

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

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

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



Languages $L$	Documents			Labels	
	$ 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.

Tagged by **journalists**

# Full-resource Scenario: Bilingual Training

		Auxiliary → English							English → Target							
		Models	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	<b>82.9</b>	69.7	86.8	80.3	79.0	76.0
		<b>MHAN-Att</b>	<b>74.0</b>	<b>74.2</b>	<b>74.1</b>	<b>72.9</b>	<b>73.9</b>	<b>73.8</b>	<b>73.3</b>	<b>72.5</b>	82.5	70.8	<b>87.7</b>	80.5	<b>82.1</b>	76.3
		MHAN-Both	72.8	71.2	70.5	65.6	71.1	68.9	69.2	70.4	82.8	<b>71.6</b>	87.5	<b>80.8</b>	79.1	<b>77.1</b>
$Y_{specific}$	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
		<b>MHAN-Att</b>	<b>46.3</b>	<b>46.0</b>	<b>45.9</b>	<b>45.6</b>	<b>46.4</b>	<b>46.4</b>	<b>46.1</b>	<b>46.5</b>	<b>46.7</b>	<b>43.3</b>	<b>47.9</b>	45.8	<b>41.3</b>	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	<b>46.1</b>	40.7	<b>50.3</b>

Input: 40-d, Encoders: Dense 100-d, Attentions: Dense 100-d Activation: 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	$\Delta\%$
en→de	0.1-0.5%	29.9	<b>41.0</b>	37.0	39.4	+37.1
	1-5%	51.3	51.7	49.7	<b>52.6</b>	+2.5
	10-50%	63.3	63.0	63.8	<b>63.8</b>	+0.7
en→es	0.1-0.5%	38.9	38.6	33.3	<b>41.5</b>	+6.7
	1-5%	45.5	50.8	<b>50.8</b>	50.1	+11.6
	10-50%	74.2	<b>75.7</b>	74.2	75.2	+2.0
en→pt	0.1-0.5%	30.9	25.3	31.6	<b>33.8</b>	+9.4
	1-5%	44.6	44.3	37.5	<b>47.3</b>	+6.0
	10-50%	60.9	61.9	62.1	<b>62.1</b>	+1.9
en→uk	0.1-0.5%	60.4	<b>62.4</b>	59.8	60.9	+3.3
	1-5%	68.2	67.7	<b>70.6</b>	69.0	+3.5
	10-50%	76.4	76.2	76.3	<b>76.7</b>	+0.3

0.5%

5%

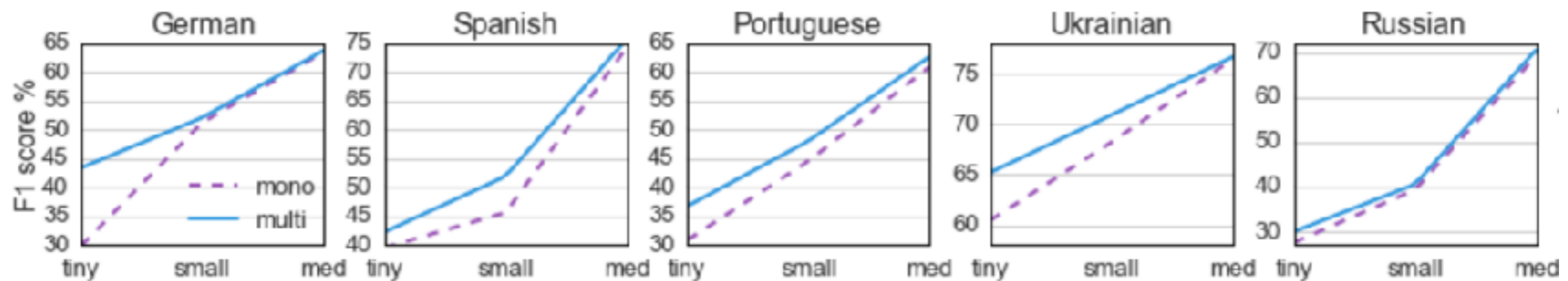
50%

Training percentage

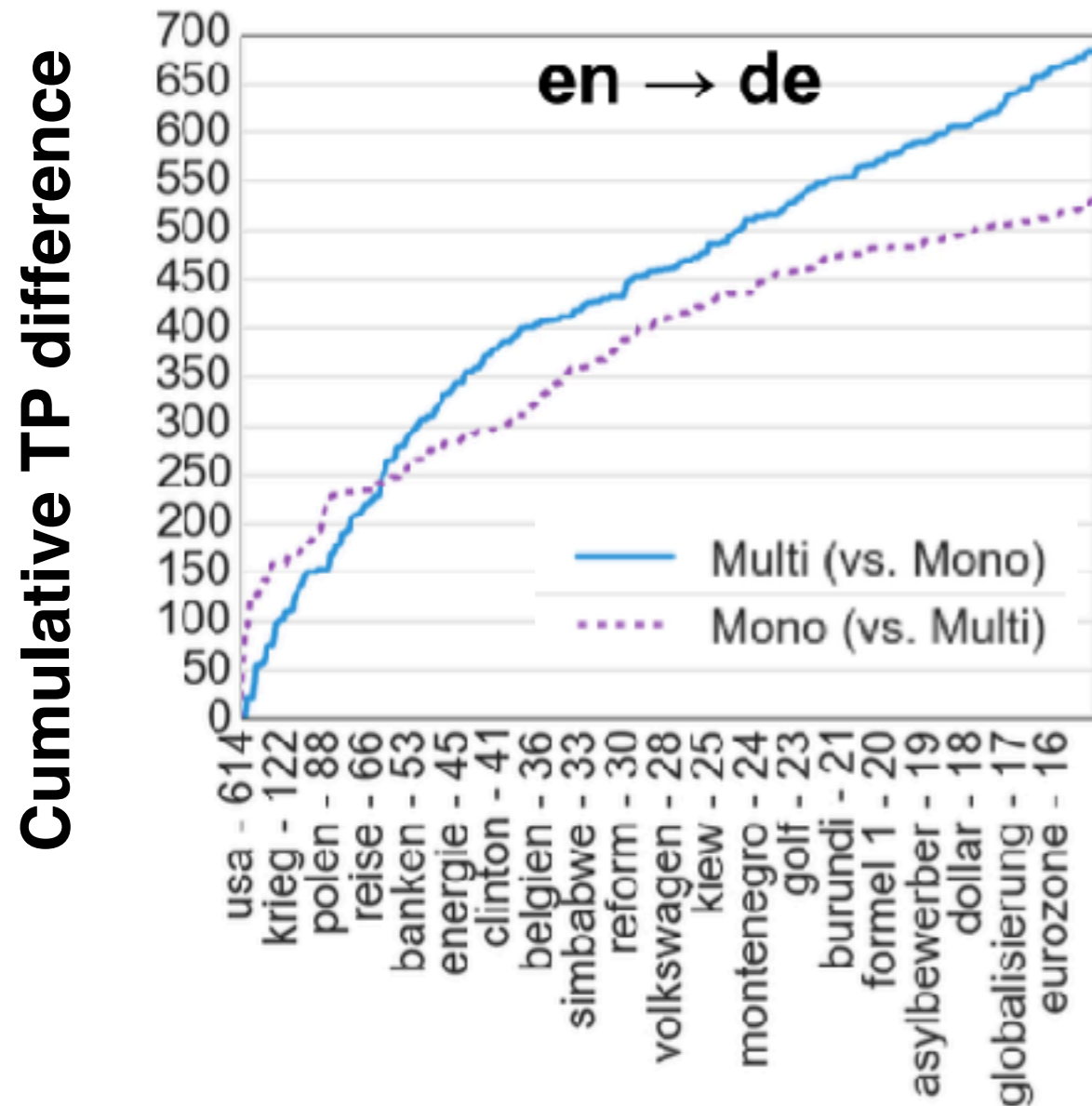
high

Improvement

low



# Qualitative Analysis: English - German



**Labels sorted by frequency**

- 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

Keyword ▼

- 0/dw\_1036083
- 1/dw\_1038974
- 2/dw\_1037081
- 3/dw\_1036545
- 4/dw\_1036547
- 5/dw\_1038196
- 6/dw\_1036286
- 7/dw\_1038178
- 8/dw\_1038047
- 9/dw\_1036618
- 10/dw\_1038181
- 11/dw\_1037461
- 12/dw\_1033672
- 13/dw\_1036551
- 14/dw\_1036544
- 15/dw\_1035770
- 16/dw\_1034899
- 17/dw\_1038555
- 18/dw\_1036299
- 19/dw\_1037051

Document (file=dw\_1037081.json)

Att. [0,1]	Document
0.023	afghanischer sonnenaufgang hat stimme ener frau
0.047	in afghanistan hat der sonnenaufgang die stimme einer frau .
0.130	zwei jahre nach dem sturz der taliban senden mit internationaler hilfe die ersten drei von frauen afghanistans - in dem zutiefst konservativen land ein mutiges unterfangen .
0.454	( sonnenaufgang ) heißt der ende oktober in betrieb gegangene sender der stadt den taliban war es verboten , musik zu hören , und frauen durften nicht arbeiten .
0.059	noch heute dürfen frauen in herat zwar im rundfunk moderieren - aber nicht singen
0.062	musikaufnahmen weiblicher interpreten sind verboten .
0.149	deswegen müssen wir bei den programmen aus kabul immer die sangerinnen herausschneiden , sagt sind neben radio sonnenaufgang noch zwei weitere von frauen betriebene radiostationen in afghanistan auf sendung .
0.002	weitere sollen in den kommenden jahren folgen .
0.002	gedacht ist vor allem daran , die landbevölkerung zu erreichen .
0.011	dort kann der rundfunk besonders nützlich sein , sagt kamal .

English  
  German  
  Spanish  
  Portuguese  
  Russian  
  Ukrainian  
  Arabic  
  Persian

English	German	Spanish	Portuguese	Russian	Ukrainian	Arabic	Persian
afghanistan (0.775)	afghanistan (0.808)	afganistán (0.606)	ataque (0.055)	ирак (0.977)	туреччина (0.256)	العراق (0.567)	کتاب رکورد کینس (0.188)
learning by ear (0.361)	taliban (0.788)	taibanes (0.507)	morte (0.042)	пресса (0.955)	росія (0.249)	أفغانستان (0.450)	بررسی صداقت (0.108)
taiban (0.353)	video (0.328)	kabul (0.171)	estados unidos (0.035)	журналист (0.878)	нато (0.229)	انفصام (0.336)	بیان اسام (0.095)
	usa (0.179)	elecciones	brasil (0.025)	печать	україна (0.202)		

# 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 Understanding of Multilingual Media



## User group meeting

July 3, 2017 Caversham, UK

Demos

Technical talks

Posters & discussions

**Contact us if interested!**

More about SUMMA:

[www.summa-project.eu](http://www.summa-project.eu)

[info@summa-project.eu](mailto: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 representations 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.