Word Representations for Named Entity Recognition

Rodrigo Agerri

HiTZ Center - Ixa

University of the Basque Country UPV/EHU

http://hitz.eus/

https://ragerri.github.io/

Contents

1. Why Named Entity Recognition?
2. Introduction to the task
3. Word Representations
4. Multilingual Language Models
    a. Issues with less-resourced languages
5. Projecting Heterogeneous Annotations

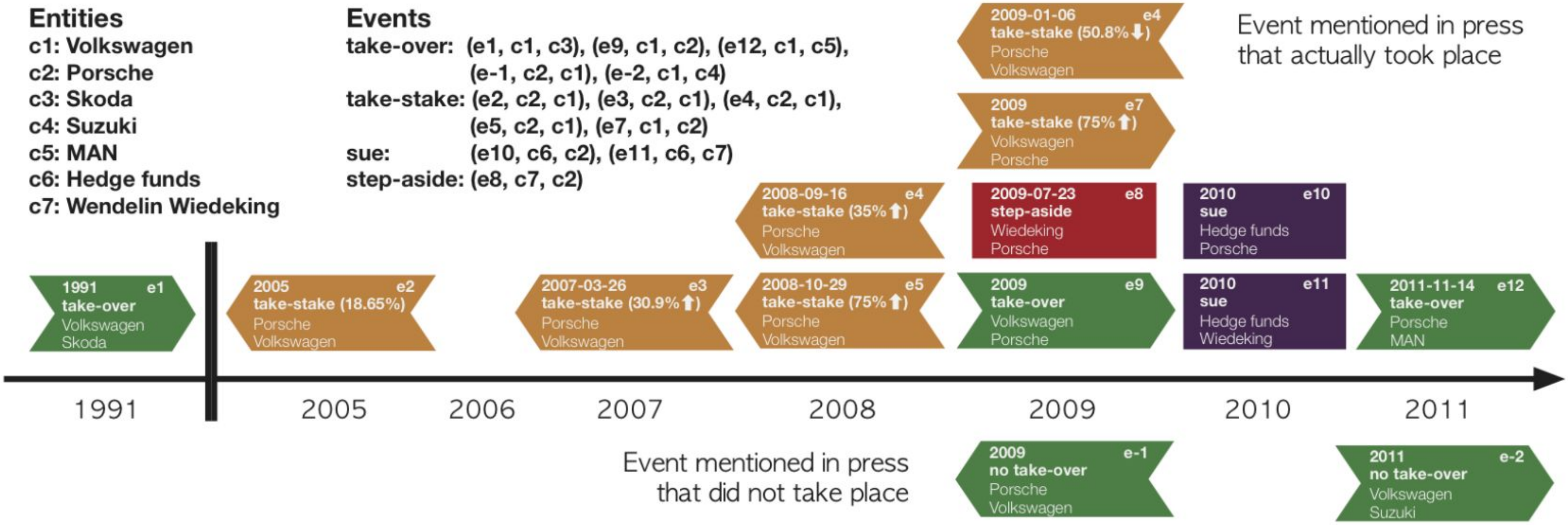
Textual Source
Following the *takeover of Škoda* Auto in 1991 by the *Volkswagen* Group
In late 2005, Porsche *took* an 18.65% *stake* in the Volkswagen Group, further *cementing their relationship*, and *preventing a takeover* of Volkswagen Group
On 26 March 2007, Porsche *took its holding* of Volkswagen AG shares to 30.9%, *triggering a takeover bid* under German Law.
Porsche could *launch a full takeover bid* for Volkswagen, Europe's biggest car manufacturer, this week if the EU's highest court makes its widely expected decision to ban a post-war law *giving* the German state *effective control over* VW.
On 16 September 2008, Porsche *increased its shares* by another 4.89%, in effect *taking control of* the company, with more than 35% of the voting rights.
Hedge funds *face* Volkswagen *storm* as Porsche *takeover* boosts shares. VW shares have *risen sharply* this week as Porsche built a 75 per cent stake, and unveiled plans to force through a deal *to take control of* the Golf and Polo car maker.
Porsche AG *took step closer to controlling* the much larger Volkswagen AG by upping its share holdings to 50.8% in late Monday trading.
6 Jan 2009 – Porsche has been on *a quest to takeover* VW for more than two years.
With present economic conditions shrinking Porsche's available cash, the automaker may have *to adjust or delay its plans to gain full control of* Volkswagen. In January, Porsche *raised its stake* in Volkswagen to 50.76% gaining a majority stake.
29-June-2009 Porsche Rejects VW Takeover Offer. The *power struggle* between German automakers Porsche and Volkswagen escalated on Monday with Porsche rejecting VW's *takeover bid* as unfeasible.
23 Jul 2009 – Porsche Chief Executive Wendelin Wiedeking has *stepped aside* in a sign that Volkswagen *takeover of* its local rival is almost secured.

http://www.newsreader-project.eu/

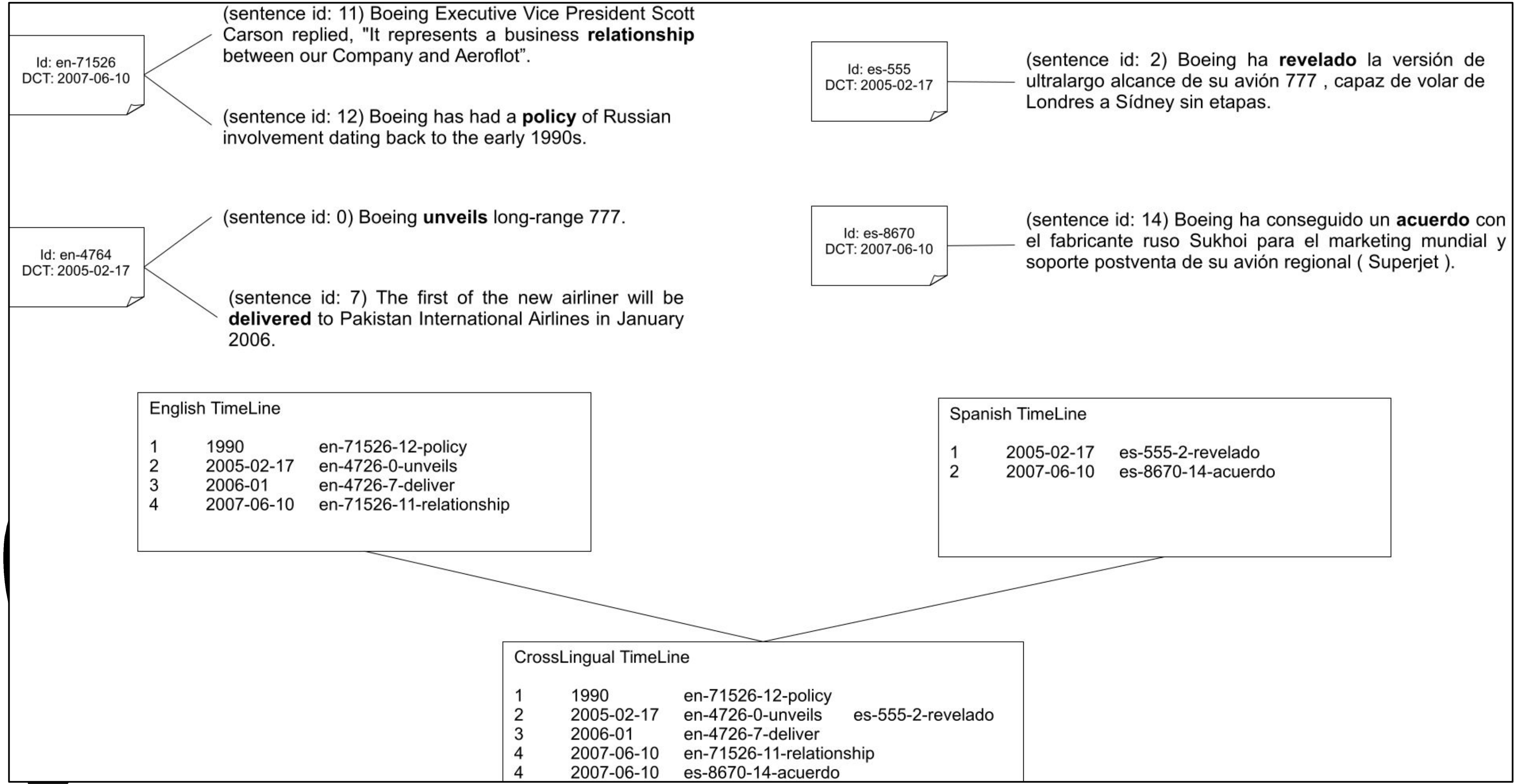
Tasks Overview

Entities
c1: Volkswagen
c2: Porsche
c3: Skoda
c4: Suzuki
c5: MAN
c6: Hedge funds
c7: Wendelin Wiedeking

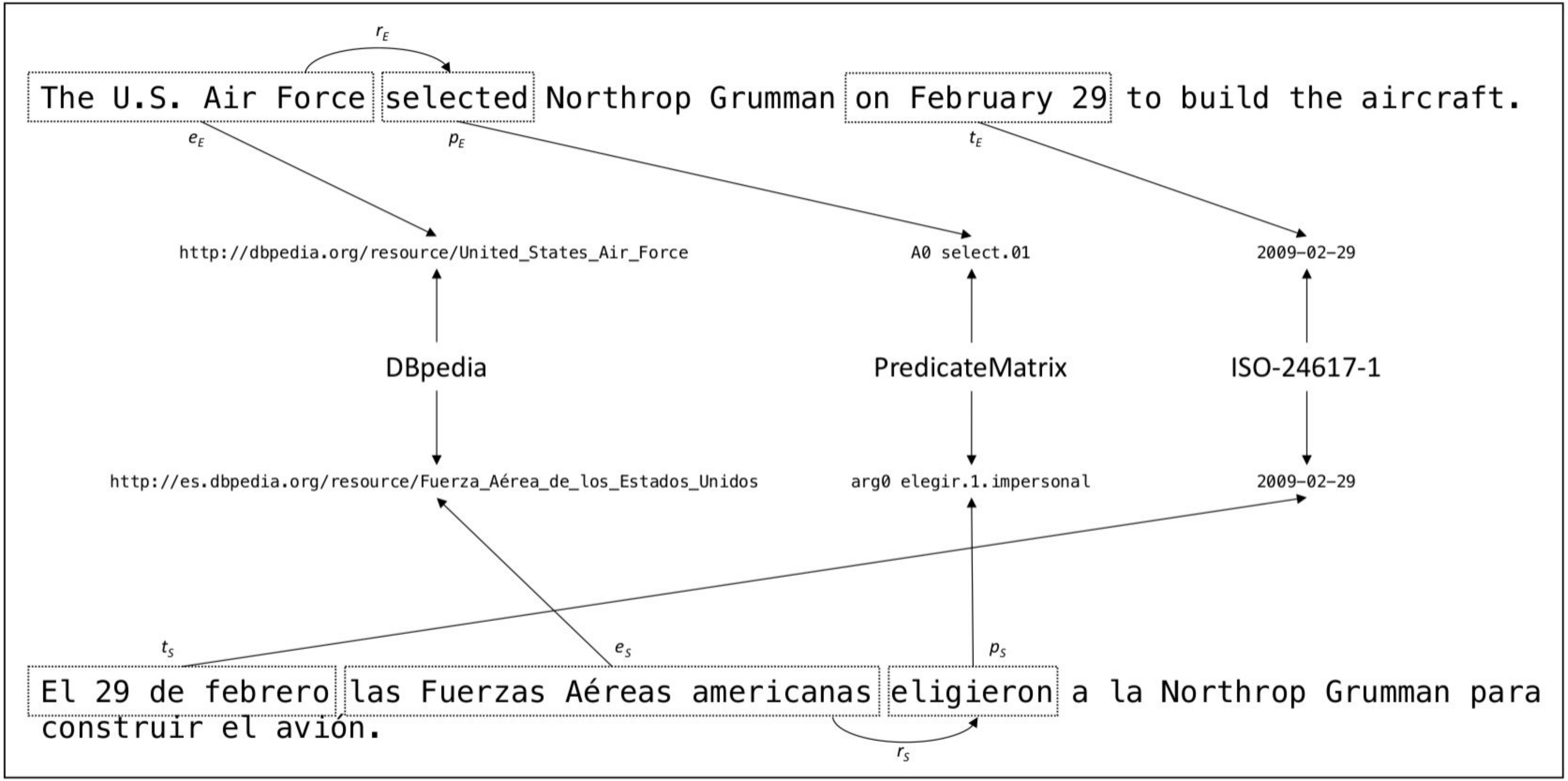
Events
take-over: (e1, c1, c3), (e9, c1, c2), (e12, c1, c5), (e-1, c2, c1), (e-2, c1, c4)
take-stake: (e2, c2, c1), (e3, c2, c1), (e4, c2, c1), (e5, c2, c1), (e7, c1, c2)
sue: (e10, c6, c2), (e11, c6, c7)
step-aside: (e8, c7, c2)



http://www.newsreader-project.eu/



Laparra et al. (2017)



Laparra et al. (2017)

Crosslingual timelines (SOTA)

		English			Spanish		
Scorer	System	P	R	F1	P	R	F1
SemEval-2015	**BTE**	24.56	4.35	7.39	12.07	4.25	6.29
	DLT	21.00	11.01	14.45	12.77	8.60	10.28
strict-evaluation	**BTE**	24.56	3.62	6.32	12.07	3.60	5.55
	DLT	21.00	9.18	12.77	12.77	7.29	9.28
relaxed-evaluation	**BTE**	24.12	5.32	8.71	11.55	5.18	7.15
	DLT	19.39	12.95	15.53	11.47	9.72	10.52

Laparra et al. (2017)

Section 2. Tasks Overview

- **Sequence Labelling:** Named Entity Recognition (NER), POS tagging, Lemmatization, Aspect Based Sentiment Analysis (ABSA), Semantic Role Labelling (SRL), Temporal Detection and Normalization
- **Document Classification:** Sentiment Analysis, Fake News, Stance, Hyper Partisanism, etc.

Contents

Named Entity Resolution

The disappearance of York University chef Claudia Lawrence is now being treated as suspected murder, North Yorkshire Police said. However detectives said they had not found any proof that the 35-year-old, who went missing on 18 March, was dead. Her father Peter Lawrence made a direct appeal to his daughter to contact him five weeks after she disappeared. His plea came at a news conference held shortly after a 10,000 reward was offered to help find Miss Lawrence. Crimestoppers said the sum they were offering was significantly higher than usual because of public interest in the case.

Named Entity Resolution (NER)

[[The disappearance of [York University chef Claudia Lawrence]] is now being treated as suspected murder, North Yorkshire Police said. However detectives said they had not found any proof that the 35-year-old, who went missing on 18 March, was dead. [Her father Peter Lawrence] made a direct appeal to his daughter to contact him five weeks after she disappeared. His plea came at a news conference held shortly after a 10,000 reward was offered to help find Miss Lawrence. Crimestoppers said [the sum] they were offering was significantly higher than usual because of public interest in the case.

Named Entity Recognition

[tim cook]**PER** is the ceo of [apple]**ORG**

Identifying spans of text that correspond to typed entities that are proper names.

BIO notation

B-PERS	I-PERS	O	O	O	O	B-ORG
tim	cook	is	the	ceo	of	apple

- **B**eginning of entity
- **I**nside entity
- **O**utside entity

[tim cook]$_{\text{PER}}$ is the ceo of [apple]$_{\text{ORG}}$

BIO notation

- Most named entity recognition datasets have flat structure (i.e., non-hierarchical labels).

✔ [The University of California]$_{\text{ORG}}$

✖ [The University of [California]$_{\text{GPE}}$]$_{\text{ORG}}$

- Mostly fine for named entities, but more problematic for general entities:

[[John]$_{\text{PER}}$'s mother]$_{\text{PER}}$ said …

Evaluation

	1	2	3	4	5	6	7
	tim	cook	is	the	CEO	of	Apple
gold	B-PER	I-PER	O	O	O	O	B-ORG
system	B-PER	O	O	O	B-PER	O	B-ORG

<start, end, type>

Precision	1/3
Recall	1/2

gold

<1,2,PER>
<7,7,ORG>

system

<1,1,PER>
<5,5,PER>
<7,7,ORG>

Learning

The classification function that we want to learn has two (main) different components:

- the formal structure of the learning method (what's the relationship between the input and output?) → Naive Bayes, logistic regression, recurrent neural network, etc.
- **the representation of the data (words?)**

Averaged Perceptron

Inputs: Training examples (x_k, y_k)
Initialization: $\overline{\lambda} = 0$
Algorithm:
For $l = 1$ to L, $k = 1$ to n
Use Viterbi to get $z_k = \operatorname{argmax}_z \overline{\lambda} \cdot \Phi(x_k, z)$
If $z_k \neq y_k$ then $\overline{\lambda} = \overline{\lambda} + \Phi(x_k, y_k) - \Phi(x_k, z_k)$
Output: weights $\overline{\lambda}$

$$\lambda_i^{av} = \sum_{l = 1 \text{ to } L,\ k = 1 \text{ to } n} \lambda_i^{l,k} / Ln$$

- *n* sentences for training
- Weights initialization = 0
- L iterations over training data
- For every labeled sentence in training, find the best sequence (z_k) using current weights
- If z_k equals to *gold sequence*, move to next sentence
- Otherwise, for every feature in the *gold* but not in prediction, add 1 to its weight, otherwise substract 1
- Average: intermediate weights assigned to every feature is divided by the number of iterations

Contents

Word representations

- One-hot representation
- Distributional Semantic Representations
- Static Word Embeddings
- Contextual Word Embeddings
  - Sub-tokens
  - Characters

One-hot representation

Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch

	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

Each word gets a 1x9 vector representation

Distributional Semantics

Distributional similarity based representations

You can get a lot of value by representing a word by means of its neighbors

"You shall know a word by the company it keeps"

(J. R. Firth 1957: 11)

One of the most successful ideas of modern statistical NLP

government debt problems turning into	banking	crises as has happened in
saying that Europe needs unified	banking	regulation to replace the hodgepodge

↖ These words will represent *banking* ↗

You can vary whether you use local or large context to get a more syntactic or semantic clustering

Adapted from Manning CS224n slides

Word Clusters

Class based models learn word classes of similar words based on distributional information (~ class HMM)

- Brown clustering (Brown et al. 1992, Liang 2005)
- Exchange clustering (Martin et al. 1998, Clark 2003)

  1. Clinton, Jiang, Bush, Wilensky, Suharto, Reagan, ...
  5. also, still, already, currently, actually, typically, ...
  6. recovery, strength, expansion, freedom, resistance, ...

Distributional representations

Cluster the words in a corpus (dimensions = clusters)

Locatives

Donostiara
Baionara
Zurichera
Gazteizera
Parisera
....

Hospitality

motel
hotel
restaurant
resort
apartment
....

Nature

mountain
hill
ridge
lake
rield
....

Corpora for cluster training

	Million words in corpus		Million words for training		
			Brown	Clark	Word2vec
en	Reuters RCV1	63	35	63	63
	Wikipedia (20141208)	1700	790	790	1700
	Gigaword 5th ed.	4000	–	–	4000
de	Wikipedia (20140725)	650	190	190	650
	deWac [6]	1100	500	500	1100
es	Wikipedia (20140810)	428	246	246	428
	elperiodico (1998–2002)	60	35	60	60
	Gigaword 3rd ed.	1150	330 (afp)	330 (afp)	1150
nl	Wikipedia (20140804)	235	128	128	235
eu	Wikipedia (20141208)	60	12	60	60
	Egunkaria (1999–2003)	38	28	38	38
	Berria (2003–2014)	90	78	90	90

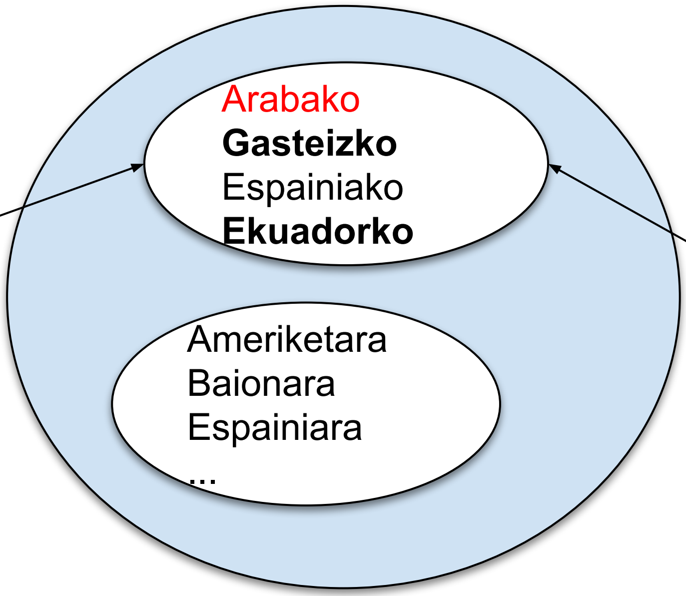
Agerri and Rigau (2016)

Clustering-based features

Training

Arabako	B-ORG
Foru	I-ORG
Aldundia	I-ORG
Arabako	B-LOC
gobernu	O
organoa	O
da.	O
Gasteizko	B-LOC
beste	O
erakunde	O
batzuekin	O
...	

Clusters



Test

Morras
Munduko
txapeldun
izan
zen
juniorretan
1994an
Ekuadorko
hiriburuan
,
Quiton.

~~training~~

Local Features

Features generated for the Basque sentence “Morras munduko txapeldun izan zen juniorretan 1994an, Ekuadorko hiriburuan, Quiton”. English: Morras was junior world champion in 1994, in the capital of Ecuador, Quito. Current token is ‘Ekuadorko’.

Feature	w_{i-2}	w_{i-1}	w_i	w_{i+1}	w_{i+2}
Token	w = 1994an	w =,	w = ekuadorko	w = hiriburuan	w =,
Token shape	wc = 1994an,4d	wc =,,other	wc = ekuadorko,ic	wc = hiriburuan,lc	wc =,,other
Previous pred	pd = null	pd = other	pd = null	pd = null	pd = other
Brown token	bt = 0111 bt = 011111		bt = 0010 bt = 001001	bt = 0101 bt = 010110	
Brown token, class	c,bt = 4d,0111 c,bt = 4d,011111		c,bt = ic,0010 c,bt = ic,001001	c,bt = lc,0101 c,bt = lc,010111	
Clark-a	ca = 158	ca = 0	ca = 175	ca = 184	ca = 0
Clark-b	cb = 149	cb = 0	cb = 176	cb = 104	cb = 0
Word2vec-a	w2va = 55	w2va = 0	w2va = 14	w2va = 14	w2va = 0
Word2vec-b	w2vb = 524	w2vb = 0	w2vb = 464	w2vb = 139	w2vb = 0
Prefix (w_i)	pre = Eku; pre = Ekua				
Suffix (w_i)	suf = o; suf = ko; suf = rko; suf = orko				
Bigram (w_i)	pw,w =,,Ekuadorko; pwc,wc = other,ic; w,nw = Ekuadorko,hiriburuan; wc,nc = ic,lc				
Trigram (w_i)	ppw,pw,w = 1994an,,,Ekuadorko; ppwc,pwc,wc = 4d,other,ic; ...				
char n-grams (w_i)	ng = adorko; ng = rko; ng = dorko; ng = ko; ng = orko ...				

Corpora used in ixa-pipe-nerc

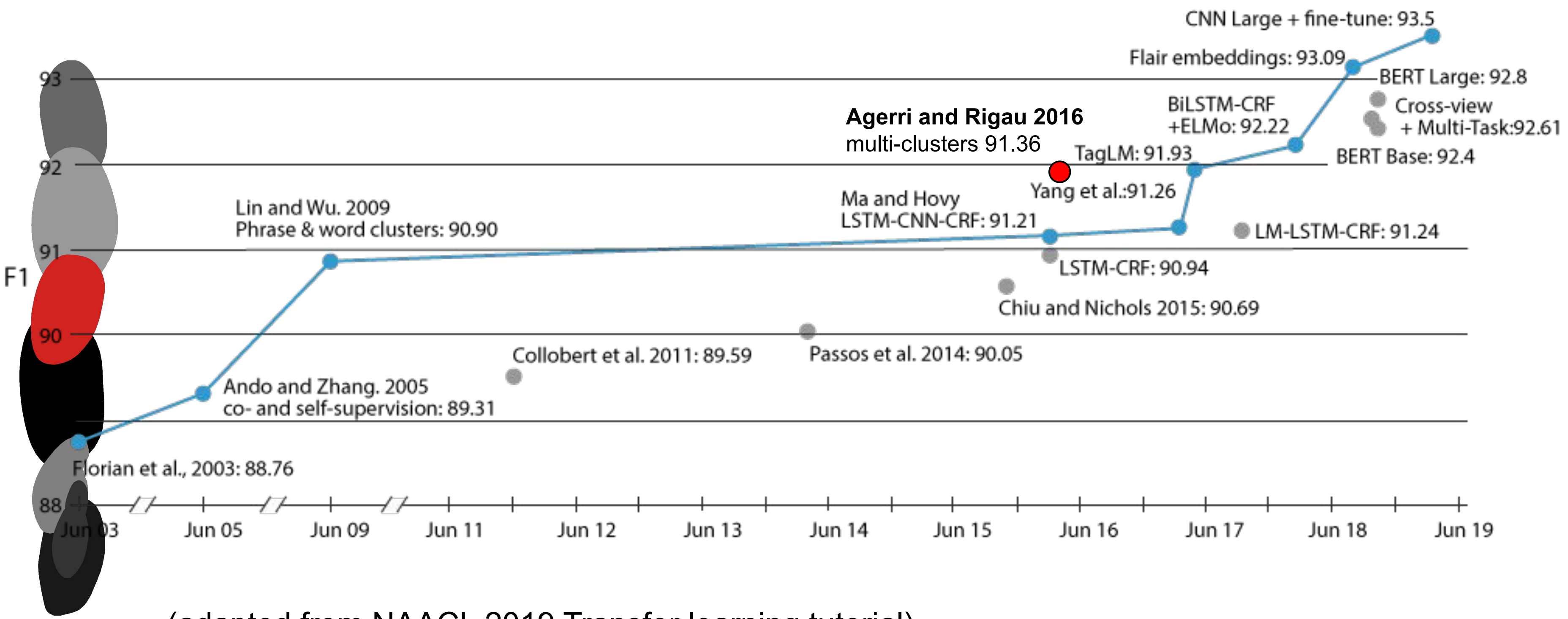
	Corpus	Source	Number of Tokens and Named Entities					
			train		dev		test	
			tok	ne	tok	ne	tok	ne
en	CoNLL 2003	Reuters RCV1	203621	23499	51362	5942	46435	5648
de	CoNLL 2003	Frankfurter Rundschau 1992	206931	11851	51444	4833	51943	3673
	GermEval 2014	Wikipedia/LCC news	452853	31545	41653	2886	96499	6893
es	CoNLL 2002	EFE 2000	264715	18798	52923	4352	51533	3558
nl	CoNLL 2002	De Morgen 2000	199069	13344	36908	2616	67473	3941
eu	Egunkaria	Egunkaria 1999-2003	44408	3817			15351	931
en	MUC7	newswire					53749	3514
	Wikigold	Wikipedia 2008					39007	3558
	Wikinews	Wikinews 2013					13957	1432
nl	SONAR-1	various genres					1000000	62505
	Wikinews	Wikinews 2013					13425	1545
es	Ancora 2.0	newswire	547198	36938				
	Wikinews	Wikinews 2013	15853	1706				

CoNLL 2003 results

	Development			Test		
Features	P	R	F1	P	R	F1
Local (L)	93.02	87.75	90.31	87.27	81.32	84.19
L + Brown reuters (BR)	92.83	89.33	91.05	90.28	86.79	88.50
L + Clark wiki 600 (CW600)	93.98	90.58	92.24	90.85	87.16	88.97
L + Word2vec giga 200 (W2VG200)	93.16	89.90	91.45	89.64	85.06	87.29
L + Word2vec wiki 400 (W2VW400)	93.22	90.02	91.59	88.98	85.09	86.99
L + BR + CW600 + W2VW400 (light)	94.16	91.96	93.04	91.20	89.36	90.27
light + CR600 + W2VG200 (comp)	94.32	92.22	93.26	91.75	89.64	90.69
comp + BW (best cluster)	94.21	92.23	93.26	91.67	89.98	**90.82**
comp + dict	**94.60**	92.78	93.68	91.86	90.53	**91.19**
BR+CR600-CW600+W2VG200+dict	94.58	92.53	93.54	**92.20**	90.19	**91.18**
charngram 1:6 + en-91-18	94.56	92.81	93.68	92.16	**90.56**	**91.36**
Stanford NER (distsim-conll03)	93.64	92.27	92.95	89.37	87.95	88.65
Illinois NER	-	-	93.50	n/a	n/a	90.57
Turian et al. (2010)	94.11	**93.81**	93.95	90.10	**90.61**	90.36
Passos et al. (2014)	-	-	**94.46**	-	-	90.90

Agerri and Rigau (2016), In Artificial Intelligence Journal.

NER (CoNLL 2003) evolution of results



(adapted from NAACL 2019 Transfer learning tutorial)

Multilingual results

NERC	eu	en	es	nl	de
ixa-pipe-nerc	75.70	91.36	84.16	85.04	76.48
Passos et al. 2014	–	90.90	–	–	–
Ratinov and Roth 2009	–	90.57	–	–	–
Stanford NER	–	88.65	–	–	–
CMP (2002-03)	–	85.00	81.39	77.05	–
C&C	–	–	–	79.63	–
Eihera	71.31	–	–	–	–
ExB (2014)	–	–	–	–	76.38

Agerri and Rigau (2016), In Artificial Intelligence Journal.

Basque results

Features	P	R	F1
Local	70.52	60.27	65.00
L + Brown egunkaria (BE)	74.54	67.59	70.90
L + Clark egunkaria 200 (CE200)	76.76	68.92	72.63
L + Clark wiki 200 (CW200)	75.57	65.60	70.23
L + Word2vec egunkaria 300 (W2VE300)	74.04	62.71	67.91
L + Word2vec berria 600 (W2WB600)	74.11	64.82	69.15
BE+C(EW)200+ W2V(E300+B600) (eu-cluster)	81.36	**73.14**	**77.03**
eu-cluster (4 classes)	**81.36**	70.78	**75.70**
Alegria et al. (2006)	72.50	70.24	71.35

Agerri and Rigau (2016), In Artificial Intelligence Journal.

Word Vector Representations

Similar idea:

Word meaning is represented as a (dense) vector – a point in a (medium-dimensional) vector space

Neural word embeddings combine vector space semantics with the prediction of probabilistic models (Bengio et al. 2003, Collobert & Weston 2008, Huang et al. 2012)

$$\textit{linguistics} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Adapted from Manning CS224n slides

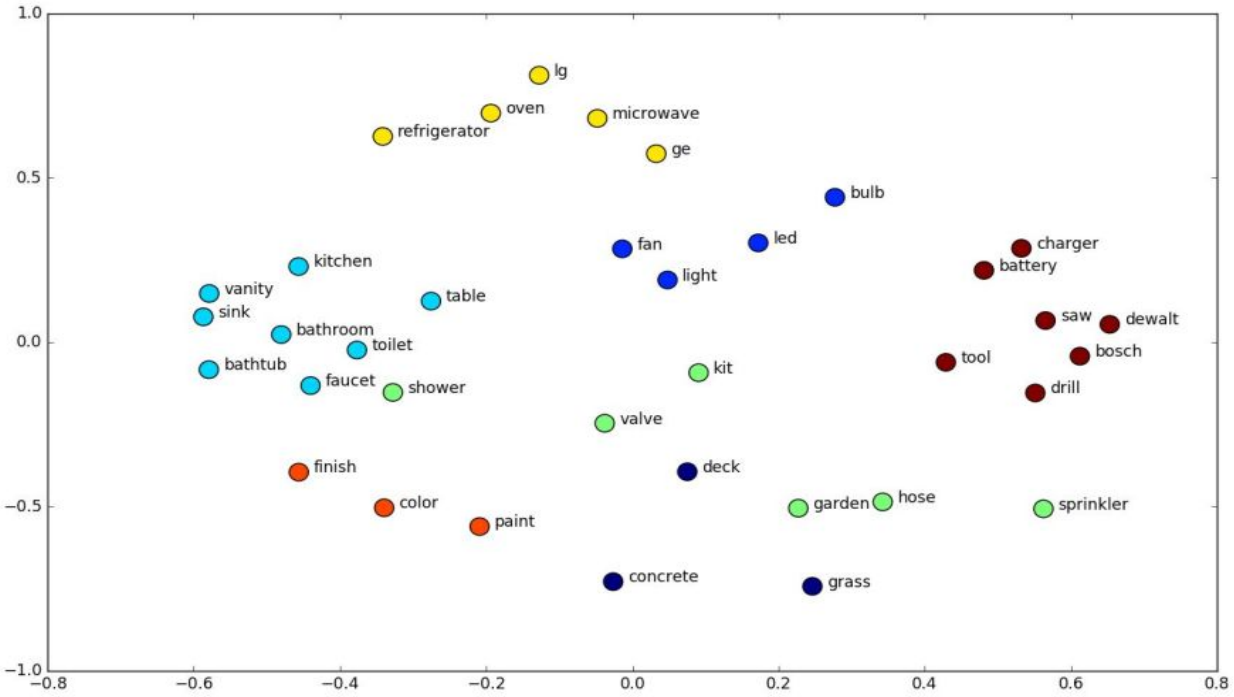
Static Word Vectors

Word vectors	Dimensions			
dog	-0.4	0.37	0.02	-0.34
cat	-0.15	-0.02	-0.23	-0.23
lion	0.19	-0.4	0.35	-0.48
tiger	-0.08	0.31	0.56	0.07
elephant	-0.04	-0.09	0.11	-0.06
cheetah	0.27	-0.28	-0.2	-0.43
monkey	-0.02	-0.67	-0.21	-0.48
rabbit	-0.04	-0.3	-0.18	-0.47
mouse	0.09	-0.46	-0.35	-0.24
rat	0.21	-0.48	-0.56	-0.37

- animal
- domesticated
- pet
- fluffy

https://projector.tensorflow.org/

Towards contextual vectors



One vector for each word in a fixed vocabulary

Towards contextual vectors

Problem 1: Word ambiguity

- "Washington"
  - Last name
  - State / city
  - Sports team
  - …

- Classic word embeddings conflate all meanings into single vector
- *Contextualized* embeddings?

Problem 2: Fixed vocabulary

- What is a word? Tokenizer decides?
  - "48-year-old"
  - "*Hotelzimmer*" (*hotel room*)
- Long-tailed distribution of words
  - Rare words?
  - Out of vocabulary words?
  - "coooooooool"
- Meaningful embeddings for any word?

One vector for each word in a fixed vocabulary

(adapted from Akbik et al. 2018)

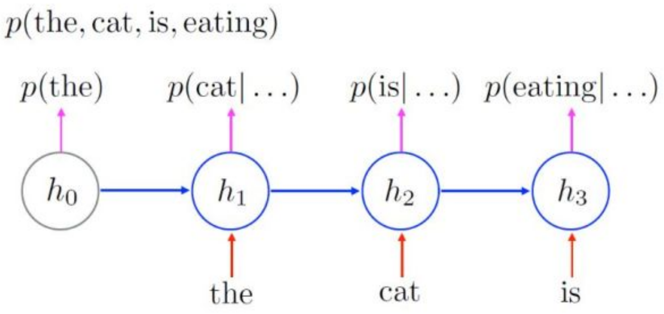
Flair contextual character-based

Language modeling:

- Train recurrent neural network (RNN) to predict the next word in a sequence of words

Character-level language modeling:

- Train RNN to predict the next *character* in a sequence of *characters*
- No tokenization
- Small vocabulary



Akbik et al. 2018. COLING

Flair character-based embeddings

what is the next word?

because it was hungry, the cat ___ **ate**

what is the next word?

because it was hungry, the cat ate ____ **the**

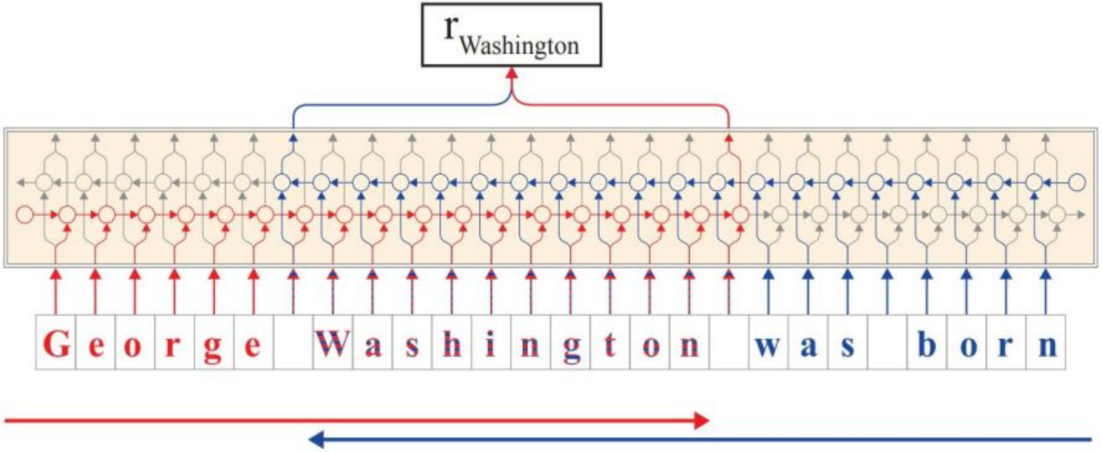
what is the next word?

because it was hungry, the cat ate the ____

The model learns

- Shallow syntax
  - nouns, verbs, adjectives
  - tense, number
- Sentence-level syntax
  - constituents
  - subordinate clauses
  - punctuation, capitalization
- Shallow semantics
  - sentiment
  - topic

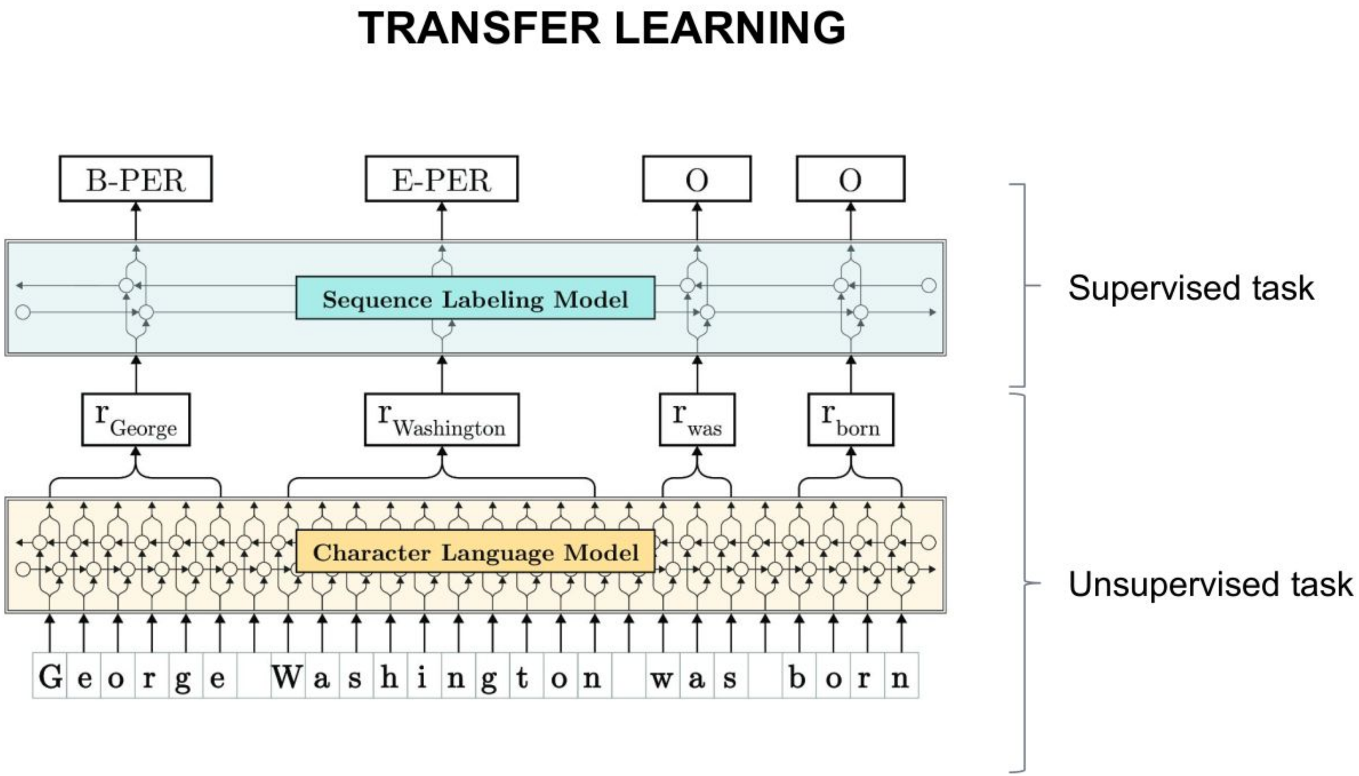
Flair: string-based and contextual



- Pass sentence as sequence of characters into two character-level language models
- Retrieve the internal states before first and after last character for each word
- Combine forward and backward states to form embedding

Akbik et al. (2018) in COLING.

Flair: string-based and contextual



Akbik et al. (2018) in COLING.

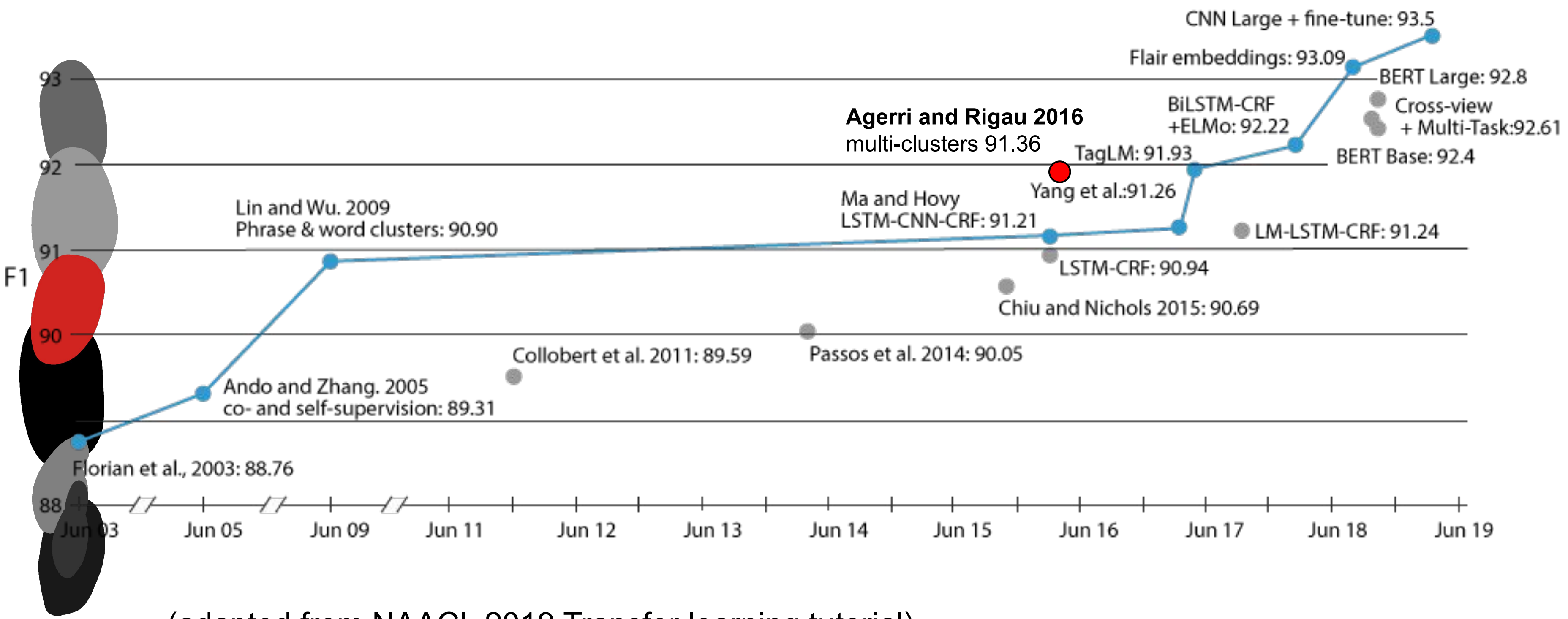
Flair: string-based and contextual

RESULTS

Approach	NER-English F1-score	NER-German F1-score	Chunking F1-score	POS Accuracy
proposed				
PROPOSED	91.97±0.04	85.78 ± 0.18	96.68±0.03	97.73±0.02
PROPOSED$_{+WORD}$	93.07±0.10	88.20 ± 0.21	96.70±0.04	97.82±0.02
PROPOSED$_{+CHAR}$	91.92±0.03	85.88 ± 0.20	**96.72**±0.05	97.8±0.01
PROPOSED$_{+WORD+CHAR}$	**93.09**±0.12	**88.32** ± 0.20	96.71±0.07	97.76±0.01
PROPOSED$_{+ALL}$	92.72±0.09	n/a	96.65±0.05	**97.85**±0.01
baselines				
HUANG	88.54±0.08	82.32 ± 0.35	95.4±0.08	96.94±0.02
LAMPLE	89.3±0.23	83.78 ± 0.39	95.34±0.06	97.02±0.03
PETERS	92.34±0.09	n/a	96.69±0.05	97.81± 0.02

Akbik et al. (2018) in COLING.

NER (CoNLL 2003) evolution of results

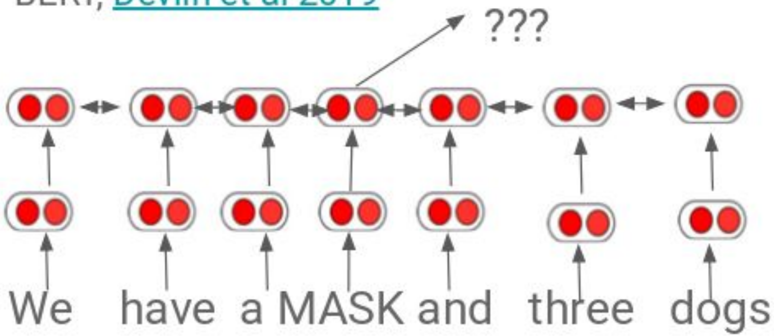


(adapted from NAACL 2019 Transfer learning tutorial)

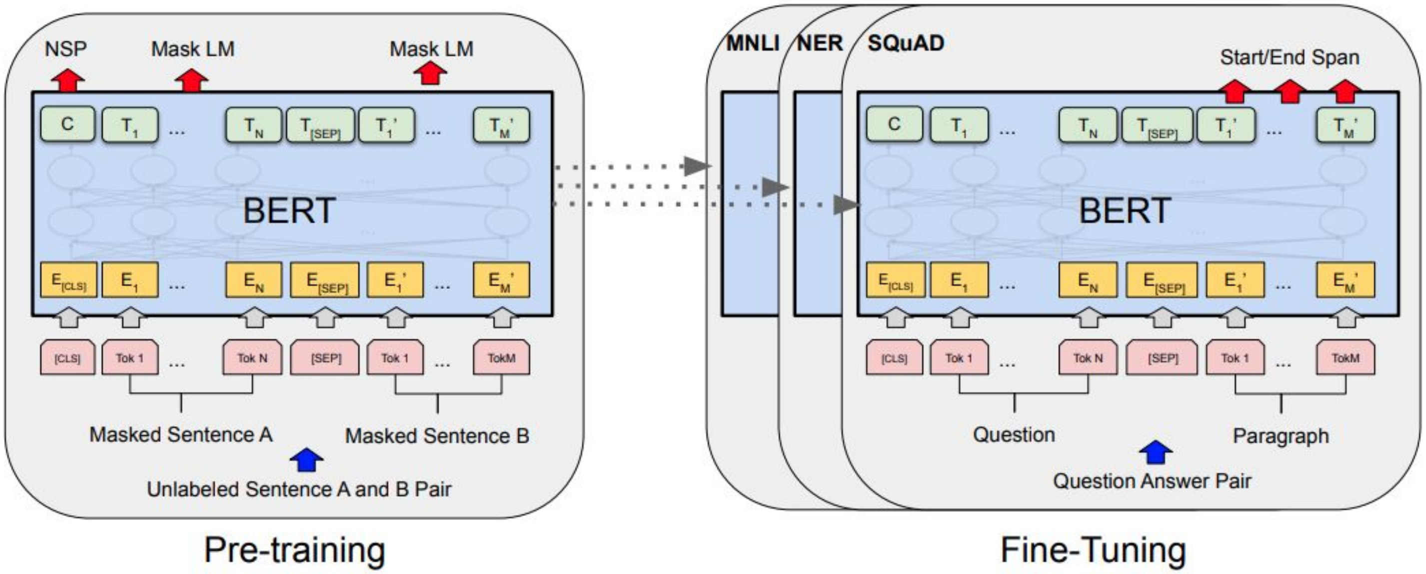
Contextual embeddings (ii)

- Instead of learning one vector per word, learn a vector that depends on context
- f(play | The kids play a game in the park.)
- f(play | The Broadway play premiered yesterday.)

BERT, [Devlin et al 2019]

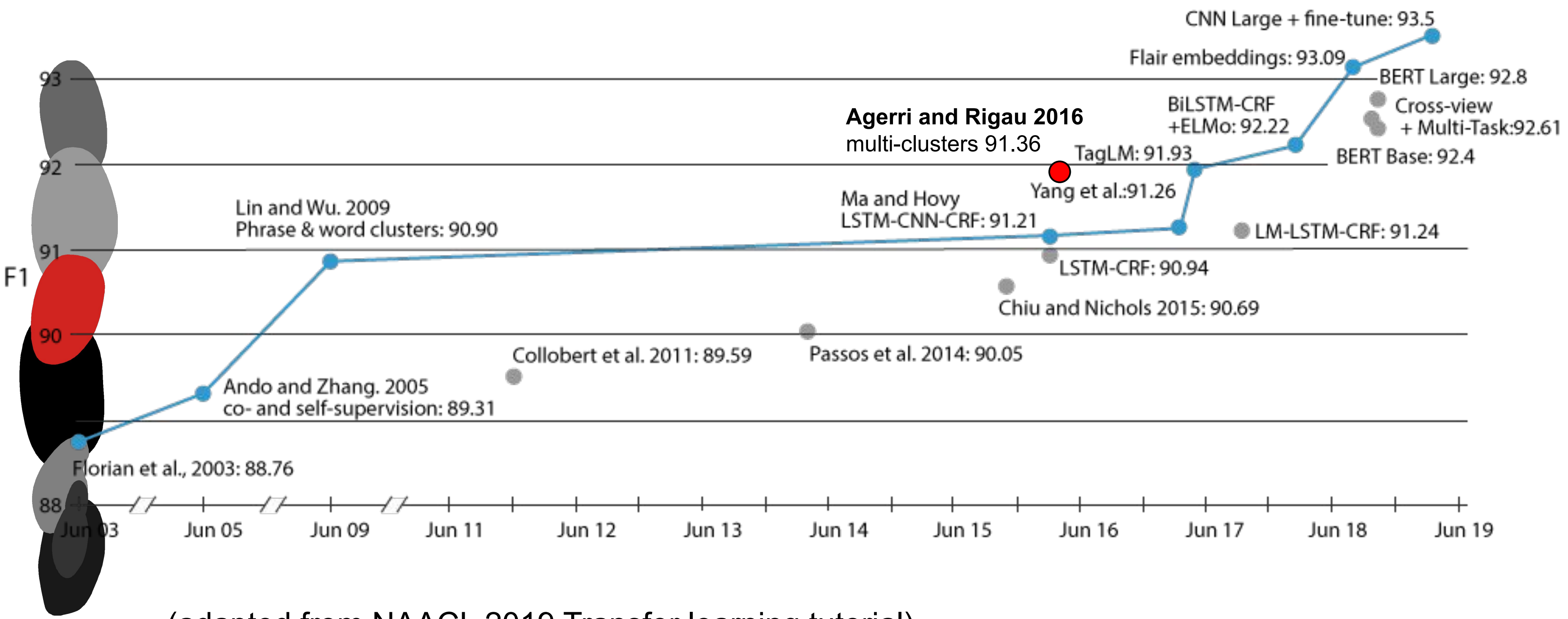


Transformers



Devlin et al 2019. In NAACL

NER (CoNLL 2003) evolution of results



(adapted from NAACL 2019 Transfer learning tutorial)

Contents

Multilingual BERT

As to why M-BERT generalizes across languages, we hypothesize that having word pieces used in all languages (numbers, URLs, etc) which have to be mapped to a shared space forces the co-occurring pieces to also be mapped to a shared space, thus spreading the effect to other word pieces, until different languages are close to a shared space.

Devlin et al . 2019. In NAACL.

Basque Morphology

Basque lemmatized	*Basque*	*Spanish lemmatized*	*Spanish*
etxe	**etxe**	**casa**	**casa**
	etxea		**casas**
	etxeak		
	etxean		
	etxearen		
	etxeek		
	etxeen		
	etxeetako		
	etxeetan		
	etxeetara		
	etxeko		
	etxekoak		
	etxera		
	etxetatik		
	etxetik		
	etxez		

Text Representations for Basque

- pre-trained language models allow to build rich multilingual representations of text (mBERT, XML-r)
- Expensive to train
- Suboptimal as less-resourced languages share the quota of substrings and parameters
- en-wiki 2.5K million words vs eu-wiki 35 million
- Tokenization
  - mBERT: Medi #kua #rene #ra
  - BERTeus: Mediku #aren #era (to-the-doctor)
    - » doctor # [the] # to

Basque Media Corpus (BMC)

	Text type	Million tokens
Wikipedia	enciclopedia	35M
Berria newspaper	news	81M
EiTB	news	28M
Argia magazine	news	16M
Local news sites	news	64.6M
BMC		224.6M

Basque NER state of the art

	Precision	Recall	F1
Static Embeddings			
FastText-Wikipedia	72.42	50.28	59.23
FastText-Common-Crawl	72.09	45.31	55.53
FastText-BMC	74.12	67.33	70.56
Flair embeddings			
Flair-official	81.86	79.89	80.82
Flair-BMC	84.32	82.66	83.48
BERT Language Models			
mBERT-official	81.24	81.80	81.52
BERTeus	87.95	86.11	**87.06**
Baseline			
(Agerri and Rigau, 2016)	80.66	73.14	76.72

Table 5: Basque NER results on EIEC corpus.

Agerri et al. (2020). In LREC

Multilingual Transformers

- Share vocabulary and representations across languages by training one model on many (100+) languages.
- Enables cross-lingual pretraining by itself
- Leads to under-representation of low-resource languages (Agerri et al. 2020)

Contents

Combine and project

- Spanish NER shared task 2020:
  - 1M corpus annotated by the Academy of the Spanish Language (RAE)

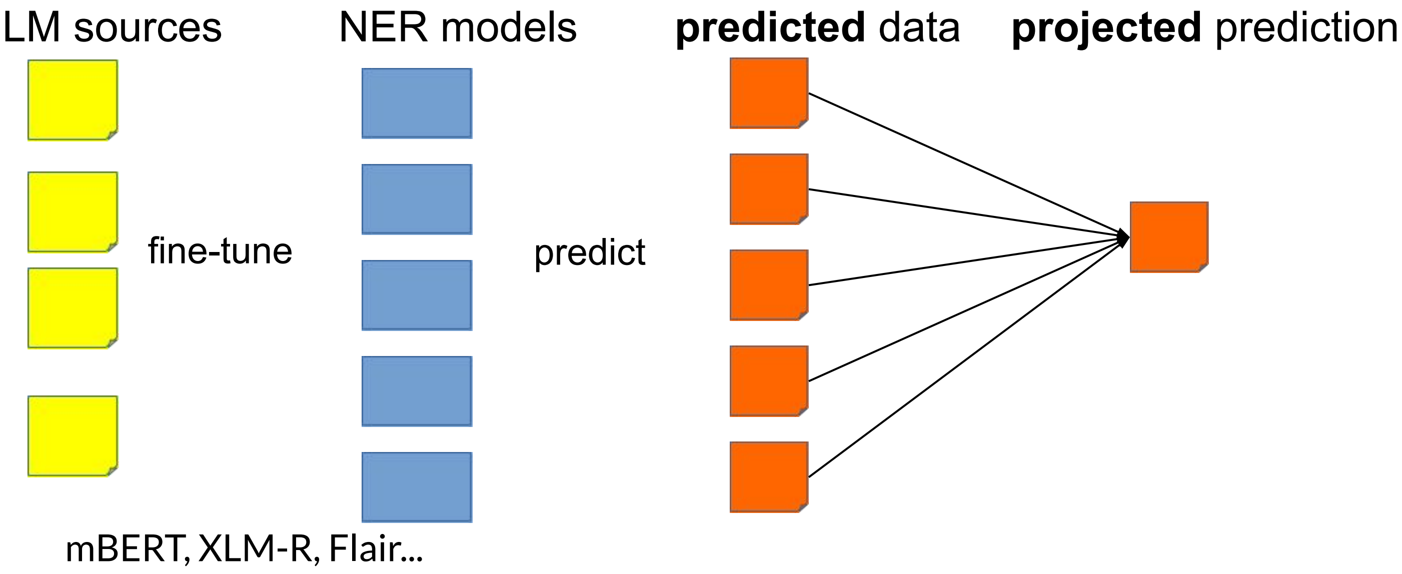
Our approach:

- **Flair LMs**: Oscar corpus, Gigaword+Wikipedia
- **Transformers**: Bertin (Gigaword+Wikipedia), XLM-RoBERTa (Common Crawl) and mBERT (Wikipedia + books)
- **Project annotations** (various strategies)

Experimental Setup

- **Flair pre-trained Spanish LM**
  - Wikipedia
- **Public pre-trained Transformer LMs:**
  - BETO (various sources)
  - XLM-RoBERTa (Common Crawl 2.5TB)
  - mBERT (Wikipedia + books)
- **Our own LMs:**
  - Flair-GW: GigaWord + Wikipedia (11GB)
  - Flair-Oscar: Oscar Spanish Corpus (157GB)
- **Project annotations**

5-1 projections



Projecting Annotations

Condition	Decision
4 > agreement	Keep Label
=< 3 agreement	Backoff:
	1. No Prediction (O) 2. Trust one system 3. Use probability scores

> 1.5 F1 score improvement over best individual system

Heterogeneity of systems/sources crucial

Results

	System	Development			Test		
		Precision	Recall	F1 score	Precision	Recall	F1 score
S1	Flair-Oscar + FT	89.65	89.36	89.51	88.86	88.63	88.74
S2	Flair-Oscar + FT (dev)	89.67	89.53	89.60	88.97	88.75	88.86
S3	Pool-Oscar + FT (dev)	**89.85**	89.63	**89.79**	89.07	**88.85**	88.96
S4	Pool-Oscar + FT e1	89.78	**89.72**	89.75	**89.29**	88.82	**89.07**
S5	Flair-Oscar + FT BIO	89.71	89.58	89.64	89.19	88.78	88.99
S6	BETO	89.64	89.34	88.99	87.19	88.36	87.77
S7	mBERT	87.90	88.90	88.40	87.03	87.75	87.39
S8	XLM-RoBERTa	88.29	89.54	88.91	87.37	88.48	87.92
P1	S2-S3-S6-S7-S8	**91.32**	**90.77**	**91.04**	90.70	88.11	89.38
P2	S2-S4-S6-S7-S8	91.10	90.59	90.84	**90.81**	88.06	89.42
P3	**S3-S4-S6-S7-S8**	91.19	90.72	90.96	90.50	**90.17**	**90.34**

> 1.3 F1 score improvement over best individual system

Capitel 2020 official results

		PER			LOC			ORG			OTH			Micro avg.			Macro avg.		
Sys	Team	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
(1)	ragerri	96.40	97.46	96.93	90.47	91.74	91.10	88.63	87.31	87.96	83.36	80.68	82.00	90.50	90.17	90.34	90.43	90.17	90.30
(2)	ragerri	96.50	97.46	96.98	90.19	91.27	90.73	88.05	87.21	87.63	84.37	81.02	82.66	90.46	90.09	90.27	90.39	90.09	90.23
(3)	ragerri	96.69	97.60	97.14	90.56	91.14	90.85	88.03	87.24	87.63	83.39	80.56	81.95	90.42	90.04	90.23	90.36	90.04	90.19
(4)	mcuadros	93.48	96.70	95.06	89.36	88.03	88.69	85.76	85.87	85.82	79.63	77.34	78.47	87.88	88.09	87.99	87.81	88.09	87.94
(5)	yanghao	94.30	96.16	95.22	87.30	89.86	88.56	84.99	85.94	85.46	79.52	77.69	78.59	87.38	88.43	87.90	87.32	88.43	87.87
(6)	lirondos	92.48	94.46	93.46	83.42	86.97	85.15	83.76	80.43	82.06	75.03	69.12	71.95	84.93	84.12	84.52	84.75	84.12	84.39
(7)	LolaZarra	91.52	92.62	92.07	83.39	80.41	81.87	80.10	83.39	81.71	78.31	73.72	75.95	83.93	83.77	83.85	83.90	83.77	83.80
(8)	lirondos	94.37	90.72	92.51	85.68	83.35	84.50	84.20	78.14	81.06	65.47	71.08	68.16	83.93	81.82	82.86	84.33	81.82	83.01
(9)	lirondos	93.23	90.09	91.63	82.05	82.54	82.29	84.55	73.85	78.84	63.89	67.17	65.49	82.67	79.46	81.03	83.01	79.46	81.11

Mila esker!
Thanks!

References

- Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa and Eneko Agirre (2020). Give your Text Representation Models some Love: the Case for Basque. In LREC 2020.
- Rodrigo Agerri, German Rigau, Language independent sequence labelling for Opinion Target Extraction. Artificial Intelligence, 268 (2019) 85-95.
- Egoitz Laparra, Rodrigo Agerri, Itziar Aldabe, German Rigau. Multi-lingual and Cross-lingual timeline extraction. Knowledge-Based Systems, 133, 77-89, 2017
- R. Agerri, G. Rigau, Robust multilingual Named Entity Recognition with shallow semi-supervised features. Artificial Intelligence, 238 (2016) 63-82.