

Information Extraction @ RALI: the Case for Named-Entity Recognition

Philippe Langlais

Huawei

October 3rd, 2019



Agenda

- RALI
- Motivation: Open Information Extraction
- How robust is NER today?
- Better NER ?
 - Learning dedicated representations with **distant supervision**
 - Better sequence labelling with **multi-tasking**
- Conclusion

RALI (Recherche Appliquée en Linguistique Informatique)



Guy Lapalme

Text Generation
Question Answering
Rule-based inspired



Jian-Yun Nie

Information Retrieval
Crosslingual Apps
Social Media Mining



Philippe Langlais

Tools for Translation
Information Extraction
Analogical Learning



**!!! We are recruiting for
September 2020 !!!**



Some Deep-Related Projects at RALI



Louis van Beurden (MSc)
Translating Weather Alerts

Environment Canada

- translation memory >> NMT > SMT
- NMT < SMT on outdomain alerts



Francis Grégoire (MSc)
Recognizing Parallel Sentence Pairs

- Siamese Network > feature-based classifier



Shivendra Bhardwaj (MSc)
Cleaning Translation Memory (MSc)

Translation Bureau

- cleaning improves NMT
- XLM > FairSeq



Guillaume Le Berre (Phd)
Inference in DL for QA

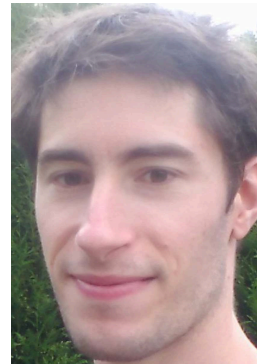
Cotutelle Univ. Lorraine (France)

- Attention on facts does not improve a BERT solution



Zakaria Soliman (MSc)
Career Path Prediction

- DL < feature-based predictors
- Issue with LinkedIn Data



Vincent Letard (PostDoc)
Entity Linking

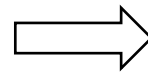
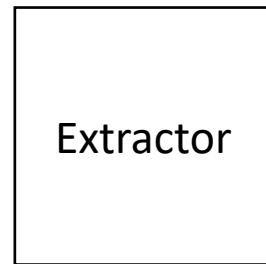
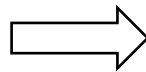
- BERT is good
- So are rules

Agenda

- RALI
- **Motivation: Open Information Extraction**
- How robust is NER today?
- Better NER ?
 - Learning dedicated representations with **distant supervision**
 - Better sequence labelling with **multi-tasking**
- Conclusion

Open Information Extraction (OIE)

- Traditional IE: narrow domains and pre-defined needs
- OIE: reads unstructured text and extracts information *tuples* without supervision [Banko et al., 2007]



| <i>arg1</i> | <i>relation</i> | <i>arg2</i> |
|--------------------|--------------------|---------------------|
| <i>Saul Hudson</i> | <i>was born in</i> | <i>Hampstead</i> |
| <i>Obama</i> | <i>declares</i> | <i>an emergency</i> |
| <i>his father</i> | <i>worked with</i> | <i>the producer</i> |

Knowledge base

Barcelona's surrender at the hands of the Nationalist forces of General Francisco Franco

- RegExps on POS tags
 - **TextRunner** [Banko et al., 2007], **ReVerb** [Fader et al., 2011], **Sonex**, etc.
- Rules applied to sentence parse tree
 - **Ollie** [Maussam et al., 2012], **ClausIE** [Del Corro and Gemulla, 2013], **MinIE** [Gashteovski et al., 2017], **Graphene** [Cetto et al., 2018], etc.
- Hybrid systems
 - **OpenIE**, etc.

REVERB

∅

OLLIE

∅

DISTYLIUM

Barcelona be surrender the hands of the Nationalist forces of General Francisco Franco 17,87

CLAUSIE

Barcelona has surrender at the hands of the Nationalist forces of General Francisco Franco -106,88

MINIE

Francisco Franco is General 0,00

Barcelona has surrender at hands of Nationalist forces of Francisco Franco 0,00

STANFORD

Barcelona has surrender at hands of Nationalist forces of General Francisco Franco 1,00

OPENIE

∅

PROPS

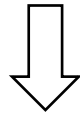
∅

No or arguably deficient triples

Dystilium

- An extension to a rule system for OIE
 - Data-driven
 - Weakly supervised
 - Able to capture non-verbal relations
 - Capable of **rephrasing tuples**

Barcelona's surrender at the hands of the Nationalist forces of General Francisco Franco



(Barcelona, fell to, Franco)

One of the many ways to leverage NER in OIE

① Relation selection

fall to

② Relation lookup in a large corpus

France *fall to* the Axis Powers in 1940 .

Again , the Broncos *fall to* the 49ers .

When Constantinople *fall to* the Turks , ...

③ Find sentences with pairs of NEs

Thing would head back downhill when
Constantinople *be take by* the Turks in 1453

In the ruin of Constantinople *'s defeat by*
the Turks later in the century .

④ Gather candidate patterns

NE₁ be take by NE₂

NE₁ 's defeat by NE₂

NE₁ under NE₂

NE₁ declare war on NE₂

...

⑤ Filter and order patterns

NE₁ be conquer by NE₂

NE₁ be take by NE₂

NE₁ be capture by NE₂

NE₁ surrender to NE₂

Agenda

- Motivation
- **Named-Entity Recognition**
 - How robust is NER today?
 - Using existing systems
 - Using proposed models
 - Better NER ?
 - Learning dedicated representations with **distant supervision**
 - Better sequence labelling with **multi-tasking**
- Conclusion

Named Entity Recognition

[PER Chilly Gonzales] (born [PER Jason Charles Beck]; [MISC 20 March 1972]) is a [MISC Canadian] musician who resided in [LOC Paris], [LOC France] for several years, and now lives in [LOC Cologne], [LOC Germany].....was signed to a three-album deal with [ORG Warner Music Canada] in [MISC 1995], a subsidiary of [ORG Warner Bros. Records]

The task consists in:

- Identifying **mentions**,
- Labeling them with a **predefined set of types**

Existing Toolkits

Named Entity Recognition for de-identification
Technical Report, Oct. 2018

Work conducted with IROSoft



Gabriel Bernier Colborne

Datasets

Onto



CoNLL



I2B2



WNUT



Fin



tokens

82k

24k

12k

2k

.5k

labels

18

4

23

6

4

In domain

| | CoNLL | FIN | i2b2 | ONTO | WNUT | Avg. | |
|----------|--------------|--------------|--------------|--------------|--------------|--------------|---------------|
| baseline | 59.10 | 41.18 | 14.60 | 20.69 | 0 | 11.76 | |
| CRF++ | 69.59 | 28.57 | 63.65 | 75.54 | 0.19 | 46.46 | |
| Stanford | 86.90 | 51.52 | 87.50 | 85.74 | 27.06 | 66.77 | Feature-based |
| Illinois | 90.57 | 57.71 | 87.76 | 87.21 | 31.38 | 68.78 | |
| NeuroNER | 89.86 | 48.18 | 91.31 | 89.20 | 33.96 | 71.49 | Deep |
| Spacy | 88.09 | 52.80 | 90.83 | 87.61 | 36.64 | 71.69 | |

Mention-level F-scores after mapping labels into MISC, PER, LOC et ORG

baseline: the most frequent label associated to each word in the training set

Out domain

| | CoNLL | FIN | i2b2 | ONTO | WNUT | Avg. |
|----------|--------|--------|--------|--------|--------|--------|
| baseline | -33.90 | -38.72 | -13.20 | +24.28 | +4.43 | +3.93 |
| CRF++ | -34.05 | +3.15 | -27.24 | -30.27 | +10.79 | -14.48 |
| Stanford | -21.90 | -37.00 | -65.29 | -18.17 | +3.49 | -26.80 |
| Illinois | -19.03 | -30.63 | -58.29 | -15.47 | +3.32 | -21.87 |
| NeuroNER | -19.40 | -22.93 | -54.86 | -16.23 | +3.30 | -23.01 |
| Spacy | -20.98 | -35.23 | -68.17 | -17.84 | -3.61 | -29.66 |

Gains of mention-level F-scores after mapping labels into MISC, PER, LOC et ORG

- Except on WNUT (small, test mentions unseen) we observe a significant drop
- Affects specific domains much more (FIN and i2b2)
- Recall more affected in general

In-domain + out-domain (data mixing)

| | CoNLL | FIN | i2b2 | ONTO | WNUT | Avg. |
|----------|--------|--------|--------|-------|--------|-------|
| baseline | -22.70 | -38.40 | -11.85 | -0.08 | +4.47 | +1.64 |
| CRF++ | -2.80 | +18.60 | +3.57 | +0.91 | +11.85 | +7.47 |
| Stanford | -2.02 | -9.01 | -4.75 | -0.78 | +4.56 | -1.43 |
| Illinois | -2.21 | -21.78 | -6.57 | -1.09 | +4.86 | -3.21 |
| NeuroNER | -0.08 | +3.76 | -2.38 | -0.38 | +5.90 | +0.38 |
| Spacy | -2.26 | -13.06 | -9.49 | -0.31 | -0.50 | -5.62 |

Gains of mention-level F-scores over training and tuning in-domain

NeuroNER seems to be the most robust

Training out-domain, fine tuning in-domain

| spacy | i2b2 |
|---------------|--------------|
| In-domain | 90.83 |
| In+out domain | 81.34 |
| Fine tuning | 90.96 |

Fine tuning is preferable, but the performance is close to training in-domain directly

Which Neural Model ?

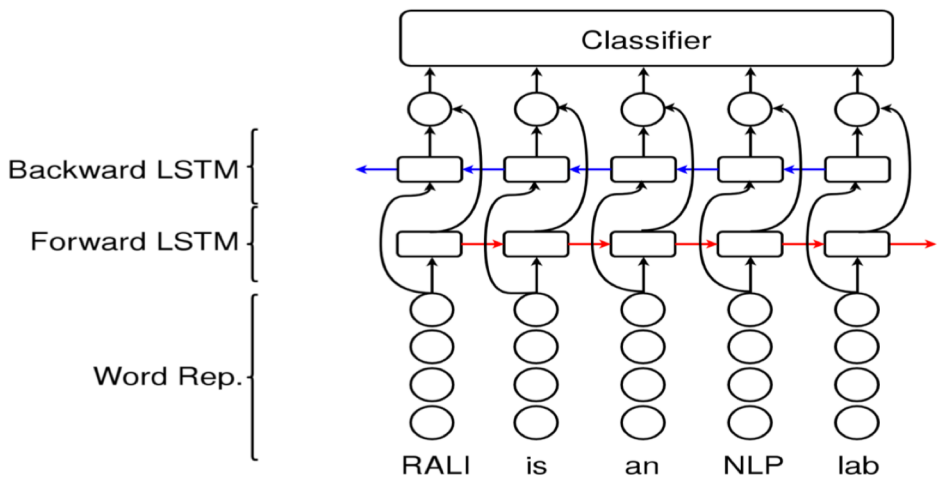
Empirical Study and Multi-task Learning Exploration
for Neural Sequence Labeling Models

[Master Thesis, Aug. 2019]

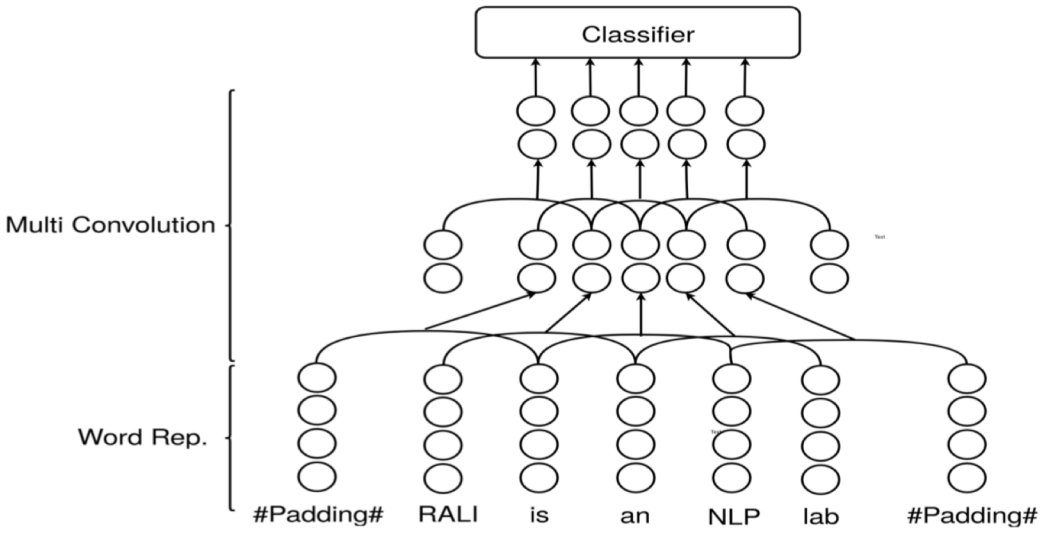


Peng Lu

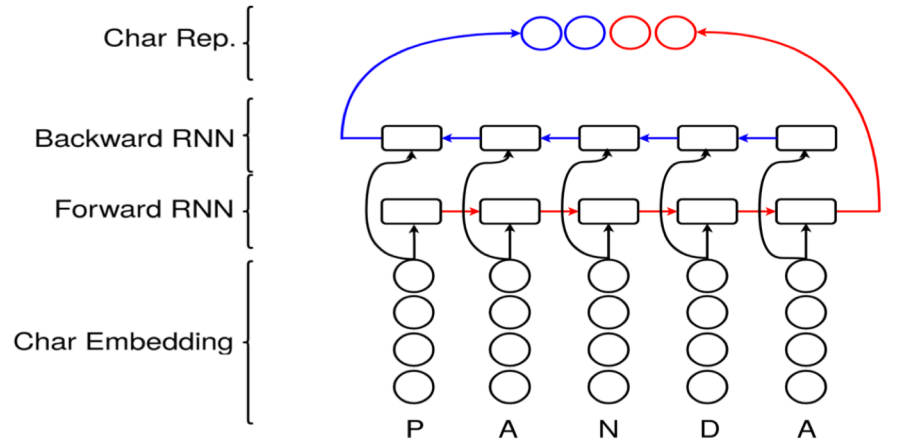
CNN/RNN, Char/Words Embeddings ...



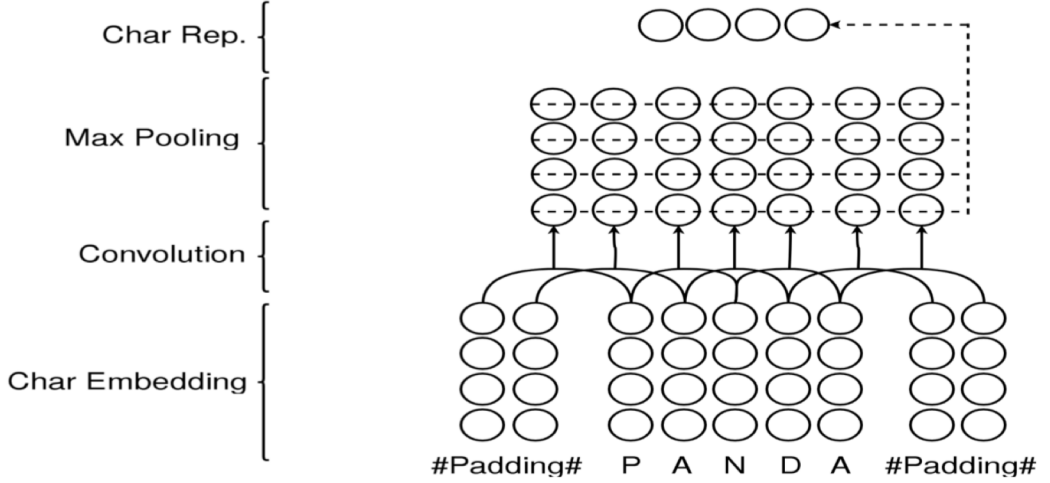
(a) Word LSTM model



(b) Word CNN model



(a) Char RNN model



(b) Char CNN model

No clear answer ...

| | No Char | Char-LSTM | Char-CNN |
|-----------------|--|---|--|
| Word-LSTM | [Ma & Hovy, 2016] [Zhai et al. 2017] [Plank et al. 2016] | [Lample et al. 2016] | [Ma & Hovy, 2016] |
| Word-CNN | [Strubell et al. 2017] | | |
| Word-LSTM + CRF | [Huang et al. 2015] | [Lample et al. 2016] [Rei, 2017] [Yasunaga et al. 2017] | [Ma & Hovy, 2016] [Chiu et Nichols, 2015] [Peters et al. 2017] |
| Word-CNN + CRF | [Collobert et al. 2012] | | [Bjerva et al. 2016] |

Controlled Comparison

- 12 models coded in the same framework (PyTorch)
- Fixed set of metaparameters
- 5 runs -> average + std

| Layer | Parameter | |
|---------------------------|-----------|-------|
| Word Embedding | type | GloVe |
| | dimension | 100 |
| Character-level Embedding | dimension | 30 |
| Dropout | rate | 0.5 |

Table 3.2. Hyper-parameters of twelve neural based models

| Layer | Parameter | |
|------------------|-------------|-----|
| Word-LSTM | hidden size | 256 |
| | layer# | 2 |
| Word-CNN | kernel size | 3 |
| | padding | 1 |
| | stride | 1 |
| | channel | 50 |
| | layer# | 4 |

| Layer | Parameter | |
|------------------|-------------|----|
| Char-LSTM | hidden size | 50 |
| | layer# | 1 |
| Char-CNN | kernel size | 3 |
| | padding | 1 |
| | stride | 1 |
| | channel | 50 |
| | layer# | 1 |

Table 3.3. Hyper-parameters of word encoding layers (left) and character encoding layers

| F1-score | | NER | | |
|-----------------|-----------------|--------------|--------------|--------------|
| | | No Char | Char-LSTM | Char-CNN |
| Word-LSTM | Reported | 87.00 [35] | 89.15 [28] | 89.36 [35] |
| | Ours (Mean±std) | 88.55 ± 0.15 | 90.63 ± 0.12 | 90.60 ± 0.22 |
| Word-CNN | Reported | 89.97 [58] | - | - |
| | Ours (Mean±std) | 88.45 ± 0.11 | 90.35 ± 0.21 | 90.33 ± 0.11 |
| Word-LSTM + CRF | Reported | 90.10 [23] | 90.94 [28] | 91.21 [35] |
| | Ours (Mean±std) | 89.98 ± 0.09 | 91.04 ± 0.10 | 91.11 ± 0.25 |
| Word-CNN + CRF | Reported | 89.59 [12] | - | - |
| | Ours (Mean±std) | 89.65 ± 0.11 | 90.47 ± 0.21 | 90.57 ± 0.11 |

Table 3.4. F1-scores on the CoNLL03 NER dataset.

- Word-LSTM > Word-CNN
- Systematic boost of the CRF layer
- Char representation always helps

Could reproduce/ouperform most of the reported results, **but** the one of [Strubell et al. 2017]

Agenda

- Motivation
- How robust is NER today?
- **Better NER ?**
 - **Learning dedicated representations with distant supervision**
 - Better sequence labelling with multi-tasking
- Conclusion

- Contextualized Word Representations from Distant Supervision with and for NER, [WNUT 2019](#)
- Robust Lexical Features For Improved Neural Network Named-Entity Recognition, [COLING 2018](#)
- Transforming Wikipedia Into A Large-Scale Fine-Grained Entity Type Corpus, [LREC 2018](#)
- WiNER: A Wikipedia Annotated Corpus For Named Entity Recognition, [IJCNLP 2017](#)



Abbas Ghaddar

Distant Supervision for NER (Nothman et al. , 2009)

Linked article texts:

Article classifications:

organisation location location

NE-tagged sentences:

[ORG Holden] is an [LOC Australian] automaker based in [LOC Port Melbourne, Victoria].

Missing links in Wikipedia

Chilly Gonzales (born **Jason Charles Beck**; 20 March 1972) is a [Canadian](#) musician who resided in [Paris](#), France for several years, and now lives in [Cologne](#), Germany. Though best known for **his** first [MC](#) and electro albums. **Gonzales** is also a pianist, producer, and songwriter..... Son was signed to a three-album deal with Warner Music Canada in 1995, a subsidiary of [Warner Bros. Records](#) While the album's production values were limited, Warner Bros. simply released the band.....

Mentions of the main entity of an article not anchored

=> detect proper nouns, noun phrases and pronouns that refer to the main entity (binary classifier) [CONLL 2016]

Missing links in Wikipedia

[Chilly Gonzales](#) (born [Jason Charles Beck](#); 20 March 1972) is a [Canadian](#) musician who resided in [Paris](#), [France](#) for several years, and now lives in [Cologne](#), [Germany](#). Though best known for [his](#) first [MC](#) and electro albums. [Gonzales](#) is also a pianist, producer, and songwriter..... [Son](#) was signed to a three-album deal with [Warner Music Canada](#) in 1995, a subsidiary of [Warner Bros. Records](#) While the album's production values were limited, Warner Bros. simply released the band.....

Wikipedians are missing links

=> following out-links

Following out-links

musician who resided in [Paris](#), **France** for several years

WIKIPEDIA The Free Encyclopedia

Paris

From Wikipedia, the free encyclopedia

Coordinates: 48°51′24″N 2°21′03″E﻿ / ﻿48.85667°N 2.35083°E﻿ / 48.85667; 2.35083

This article is about the capital of France. For other uses, see [Paris \(disambiguation\)](#).

Paris (French: [paʁi] [ⓘ]) is the capital and most populous city in **France**. It has an area of 105 square kilometres (41 square miles) and a population of 2,229,621 in 2013 with in the metropolitan area^[?]. The city is both a commune and a metropolitan area that forms the centre and **France**. The Paris Region, which has an area of 12,012 square kilometres (4,638 square miles) and a population in 2016 of 12,142,802, comprising roughly 18 percent of the population of France.^[4] By the 17th century, Paris was one of Europe's major centres of finance, commerce, fashion, science, and the arts, and it retains that position still today. The Paris Region had a GDP of €624 billion (US \$687 billion) in 2012, accounting for 30.0 percent of the GDP of France and ranking it as one of the wealthiest regions in Europe.

The City of Paris is but the core of a built-up area that

Paris
Commune and department

now lives in [Cologne](#), **Germany**

WIKIPEDIA The Free Encyclopedia

Cologne

From Wikipedia, the free encyclopedia

Coordinates: 50°56′11″N 6°57′10″E﻿ / ﻿50.93639°N 6.95278°E﻿ / 50.93639; 6.95278

This article is about the German city. For the style of perfume, see [Eau de Cologne](#). For other uses, see [Cologne \(disambiguation\)](#) and [Köln \(disambiguation\)](#).

"Koln" redirects here. It is not to be confused with [KOLN](#).

Cologne (English pronunciation: /kəˈloʊn/; German: ***Köln***, pronounced [kœln] [ⓘ]) is the largest city in the German federal State of North Rhine-Westphalia and the fourth-largest city in **Germany** (after Berlin, Hamburg, and Munich). It is

Cologne
Köln

Missing links in Wikipedia

[Chilly Gonzales](#) (born [Jason Charles Beck](#); 20 March 1972) is a [Canadian](#) musician who resided in [Paris](#), [France](#) for several years, and now lives in [Cologne](#), [Germany](#). Though best known for [his](#) first [MC](#) and electro albums. [Gonzales](#) is also a pianist, producer, and songwriter..... [Son](#) was signed to a three-album deal with [Warner Music Canada](#) in 1995, a subsidiary of [Warner Bros. Records](#) While the album's production values were limited, [Warner Bros.](#) simply released the band.....

First mention on an entity only is anchored (at best)

=> iterate !

- [Chilly Gonzales](#) → {Gonzales, Jason Charles Beck, the performer}
- [France](#) → {French Republic, Kingdom of France, the country}
- [Warner Bros. Records](#) → {[Warner Bros.](#), Warner, the company}

Statistics

| | #Links | #Documents | Links per Doc |
|---------------------------------|--------|------------|---------------|
| Wikipedia (dump 2014) | 71.5M | 4.3M | 16.6 |
| Raganato et al. 2016 | 162.6M | 4.3M | 37.8 |
| Our approach (dump 2013) | 182.7M | 3.2M | 57.0 |

95.1M proper names, 62.4M noun phrases and 24.2M pronouns

- 3% of tokens in Wikipedia natively anchored, 30% after our process
- Comes with some noise...
 - 77% correct according to an evaluation on 1000 annotations

Eldridge Pope was a traditional brewery.....Sixteen years later the Pope Brothers floated the business...

From anchored mentions to annotations

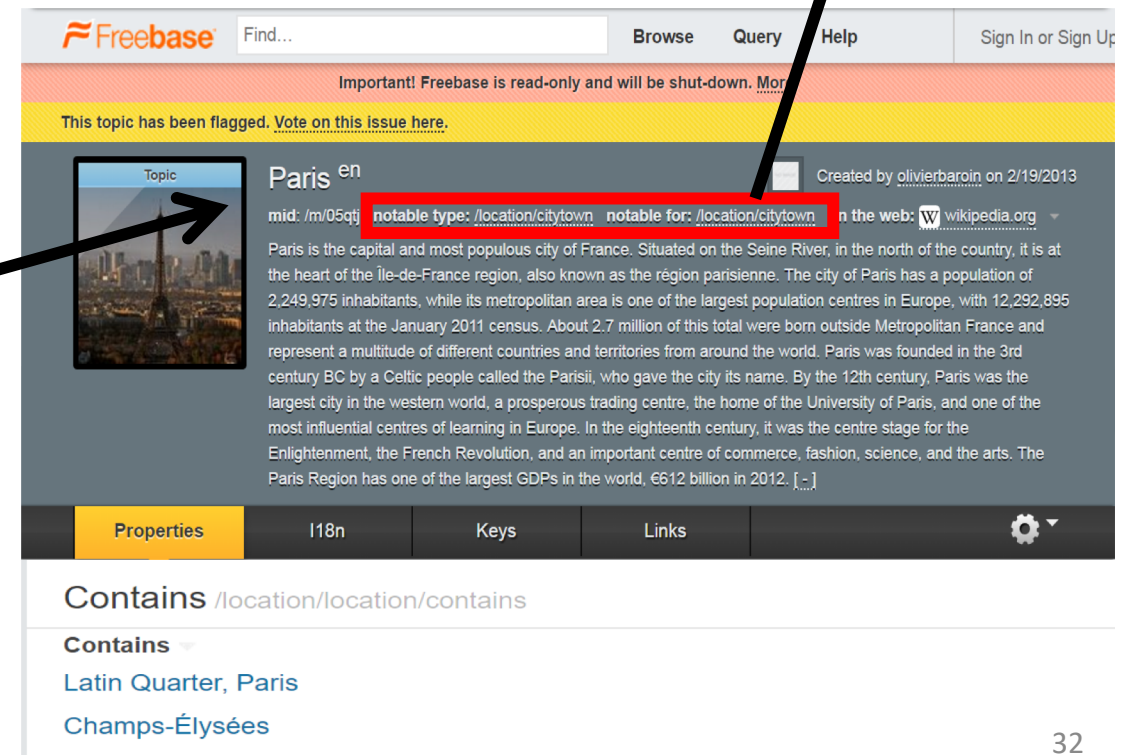
....is a [Canadian](#) musician who resided in [Paris](#)

....is a [Canadian](#) musician who resided in [LOC Paris]

notable type: [/location/citytown](#)



The screenshot shows the Wikipedia article for Paris. The title "Paris" is at the top. Below it, there's a coordinate box showing "Coordinates: 48°51'24"N 2°21'03"E". A note says "This article is about the capital of France. For other uses, see Paris (disambiguation)." The main text starts with "Paris (French: [paʁi] (listen)) is the capital and most populous city of France. It has an area of 105 square kilometres (41 square miles) and a population of 2,229,621 in 2013 within its administrative limits." There are several arrows pointing from the text boxes above to specific parts of the page: one to the "Paris" title, one to the coordinate box, and one to the "Commune and department" section header.



The screenshot shows the Freebase entry for Paris. The title "Paris en" is at the top. Below it, there's a "Topic" image of the Eiffel Tower. The main text starts with "Paris is the capital and most populous city of France. Situated on the Seine River, in the north of the country, it is at the heart of the Île-de-France region, also known as the région parisienne. The city of Paris has a population of 2,249,975 inhabitants, while its metropolitan area is one of the largest population centres in Europe, with 12,292,895 inhabitants at the January 2011 census." There are several arrows pointing from the text boxes above to specific parts of the page: one to the "Paris en" title, one to the "Topic" image, and one to the "notable type: /location/citytown" field.

4 classes: PER ORG LOC MISC

[PER Chilly Gonzales] (born [PER Jason Charles Beck]; 20 March 1972) is a [MISC Canadian] musician who resided in [LOC Paris], [LOC France] for several years, and now lives in [LOC Cologne], [LOC Germany]. Though best known for his first [MISC MC] and electro albums, [PER Gonzales] is also a pianist, producer, and songwriter..... [MISC Son] was signed to a three-album deal with [ORG Warner Music Canada] in 1995, a subsidiary of [ORG Warner Bros. Records] While the album's production values were limited, [ORG Warner Bros.] simply released the band.....

Available here: <http://rali.iro.umontreal.ca/rali/?q=en/wikipedia-main-concept>

On **October 9, 2009**, the **Norwegian Nobel Committee** announced that **Obama** had won the **2009 Nobel Peace Prize**.

On **/date**, the **/organization/government_agency** announced that **/person/politician** had won the **/award**.

(Automatically) projected into 2 fine-grained entity types:

- **FIGER**: 120 types [Lin and Weld, 2012]
 - 2-level: **/person**, **/person/musician**
- **Gillick**: 89 types [Gillick et al. 2014]
 - 3-level **/person**, **/person/artist**, **/person/artist/musician**

(Lin and Weld. 2012)

| | | | | |
|-----------------|---|--|--|--|
| person | doctor actor architect artist athlete author coach director | engineer monarch musician politician religious_leader soldier terrorist | organization | terrorist_organization government_agency government political_party educational_department military news_agency |
| location | body_of_water city country county province railway road bridge | island mountain glacier astral_body cemetery park | product | camera mobile_phone computer software game instrument weapon |
| | | | art | written_work film play music |
| | | | event | military_conflict attack election protest natural_disaster sports_event terrorist_attack |
| building | airport dam hospital hotel library power_station restaurant sports_facility theater | time color award educational_degree title law ethnicity language religion god | chemical_thing biological_thing medical_treatment disease symptom drug body_part living_thing animal food | website broadcast_network broadcast_program tv_channel currency stock_exchange algorithm programming_language transit_system transit_line |

Ambiguity

- Mapping a mention to its type is a 1-n problem
 - Several `object-type` properties in DBPedia
 - Only a few valid in a given context

Chilly Gonzales

/person

/person/artist

/person/artist/**musician**

/person/artist/**actor**

/person/artist /**author**

- 23% of mentions are ambiguous

Disambiguation

Using 2 rules:

(a) **Gonzales** was born on 20 March 1972 in **Montreal**, Canada .

person, artist, musician,
actor, auhor

rel: /people/person/place_of_birth

(b) Additionally , **he** has collaborated with **Jamie Lidell** on the albums *Multiply* and *Compass*.....

person, artist, musician,
actor, auhor

person, artist, musician

Desambiguation (evaluation)

On 1000 ambiguous mentions

| Heuristic | Pre | Rec | F1 |
|-----------------------------|-------------|--------------|-------------|
| w/o Rules | 31.8 | 100.0 | 48.3 |
| Rule-1 only | 48.8 | 87.2 | 62.3 |
| Rule-2 only | 56.4 | 85.6 | 68.0 |
| Both Rules | 79.2 | 81.8 | 80.5 |
| Level of Application | | | |
| Sentence | 66.5 | 85.5 | 73.7 |
| + Paragraph | 72.7 | 82.6 | 78.6 |
| + Section | 79.2 | 81.8 | 80.5 |

WiFiNE

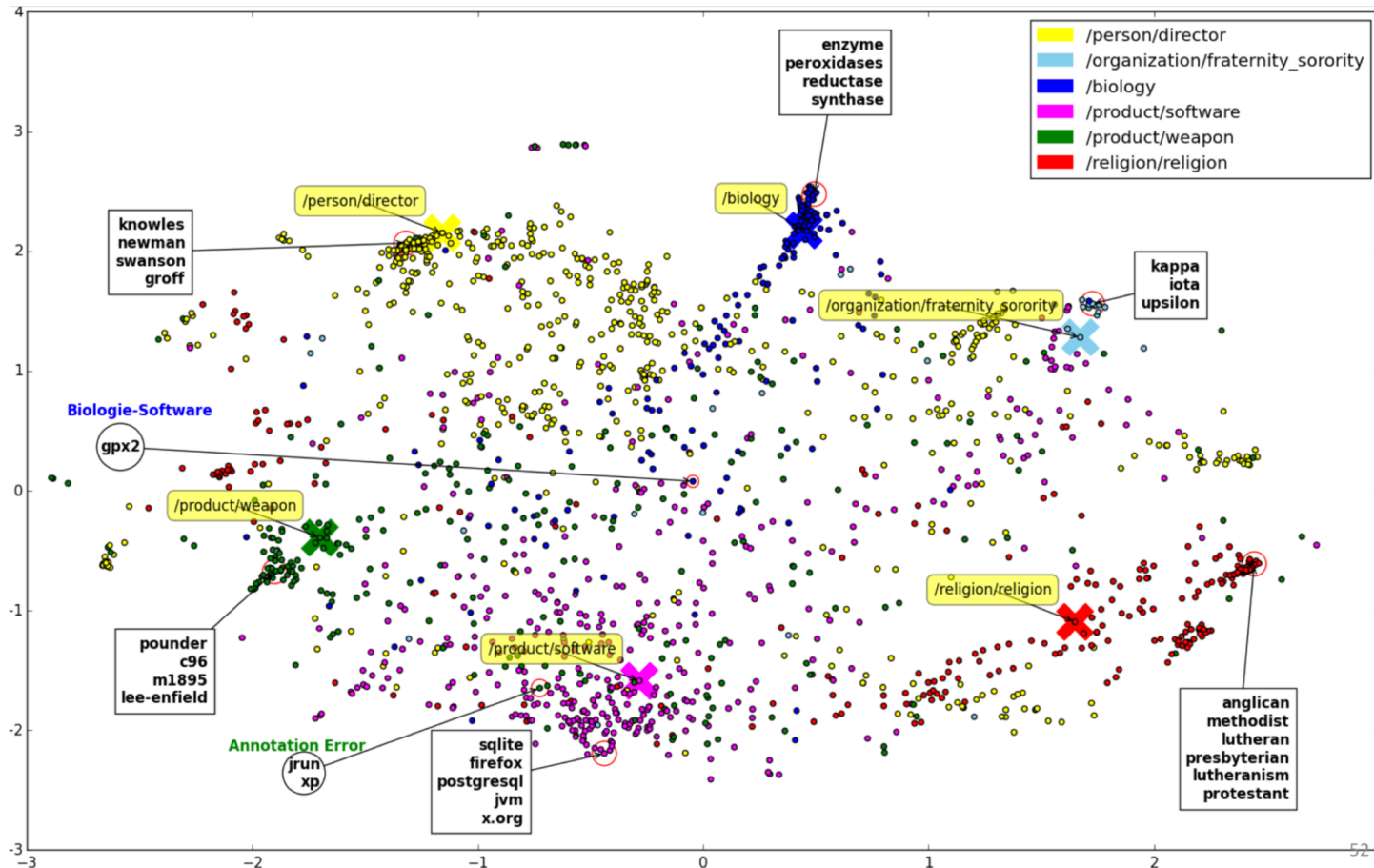
| | FIGER | GILLICK |
|----------------------------|--------------|----------------|
| Total mentions | 159.4 | 111.1 |
| Proper mentions | 82.5 (52%) | 64.8 (58%) |
| Nominal mentions | 55.9 (35%) | 29.8 (27%) |
| Pronominal mentions | 21.0 (13%) | 16.5 (15%) |
| Total Labels | 243.2 | 230.9 |
| Level 1 | 153.8 (63%) | 111.1 (48%) |
| Level 2 | 89.5 (37%) | 90.0 (39%) |
| Level 3 | - | 29.8 (13%) |

In millions

Available here: <http://rali.iro.umontreal.ca/rali/?q=en/wikipedia-main-concept>

Embedding words and labels in the same space using FastText (Bojanowski et al. 2016)

TSNE view of 6 selected types and 1500 randomly sampled single-word mentions labelled with these types in WiFiNE

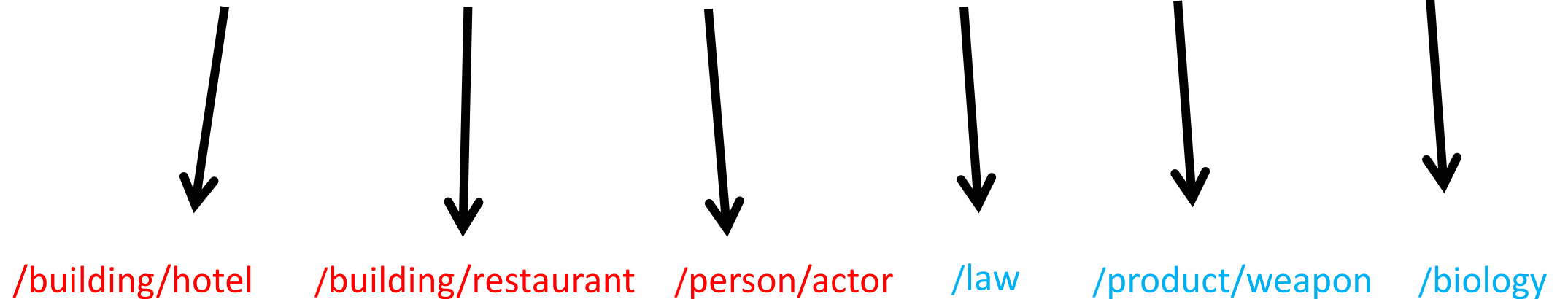


[COLING 2018]

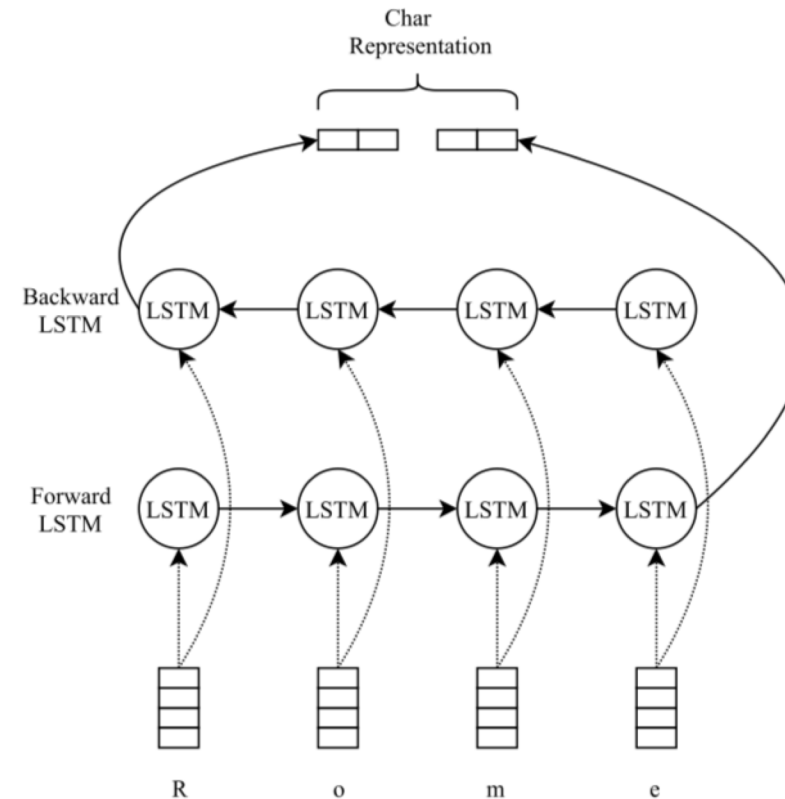
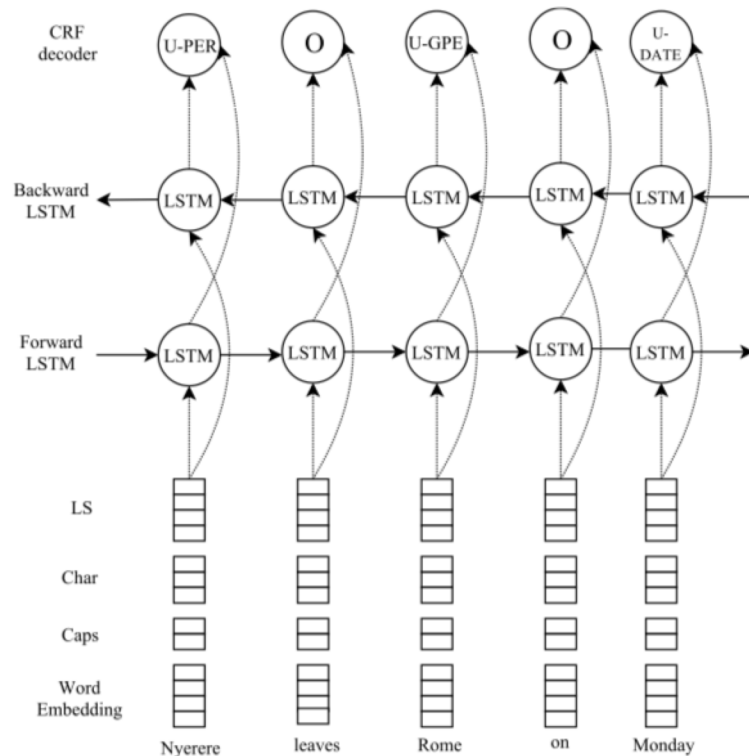
Word = vector (d=120)

Each word represented by its cosine similarity to each vector associated to type

Hilton = [..., 0.59, ..., 0.47, ..., 0.36,, 0.092, ..., 0.08, ..., 0.06, ...]

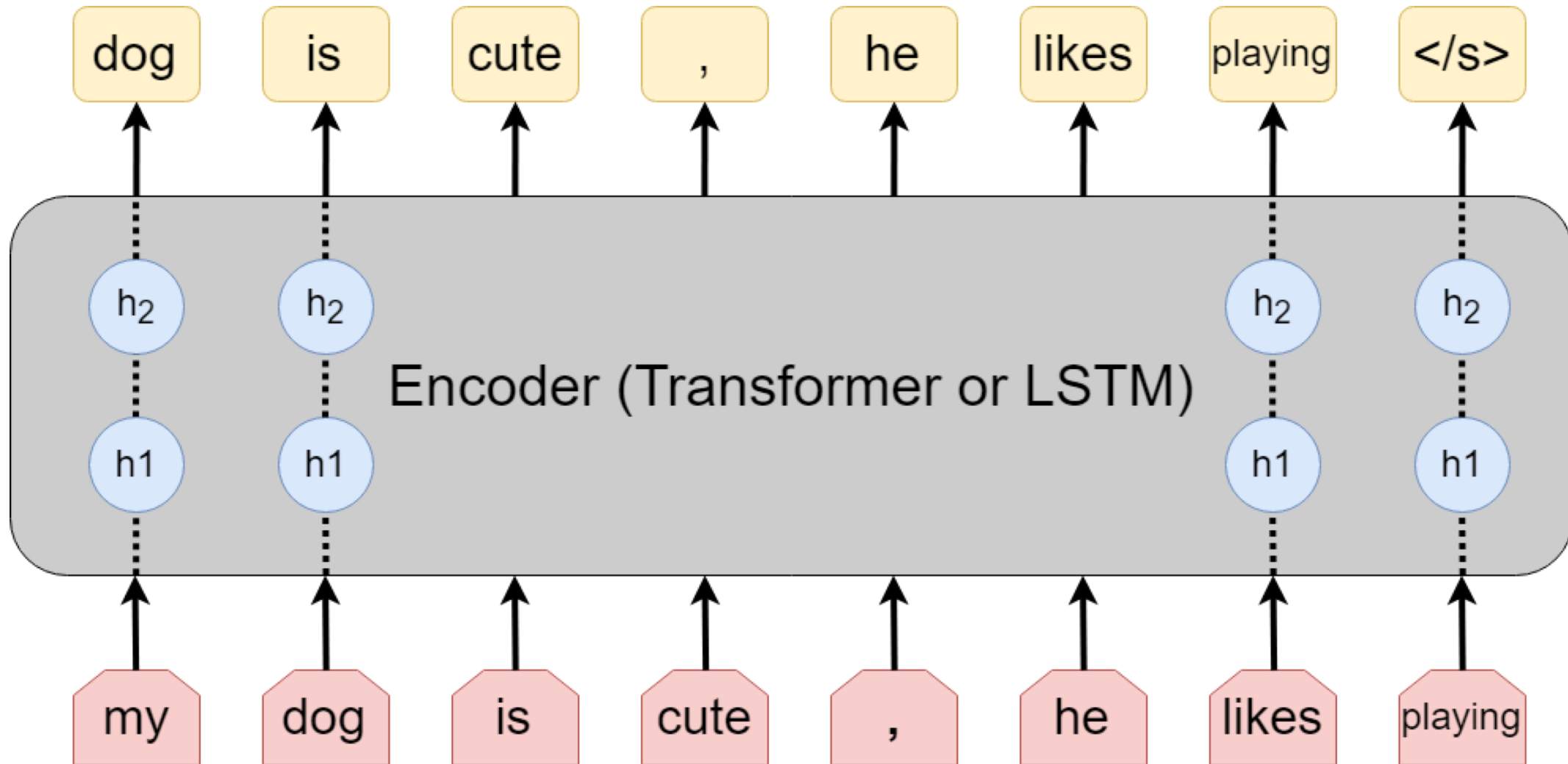


BiLSTM-CRF (Lample et al 2016; Chiu and Nichols, 2016 ; Søgaard and Goldberg, 2016)

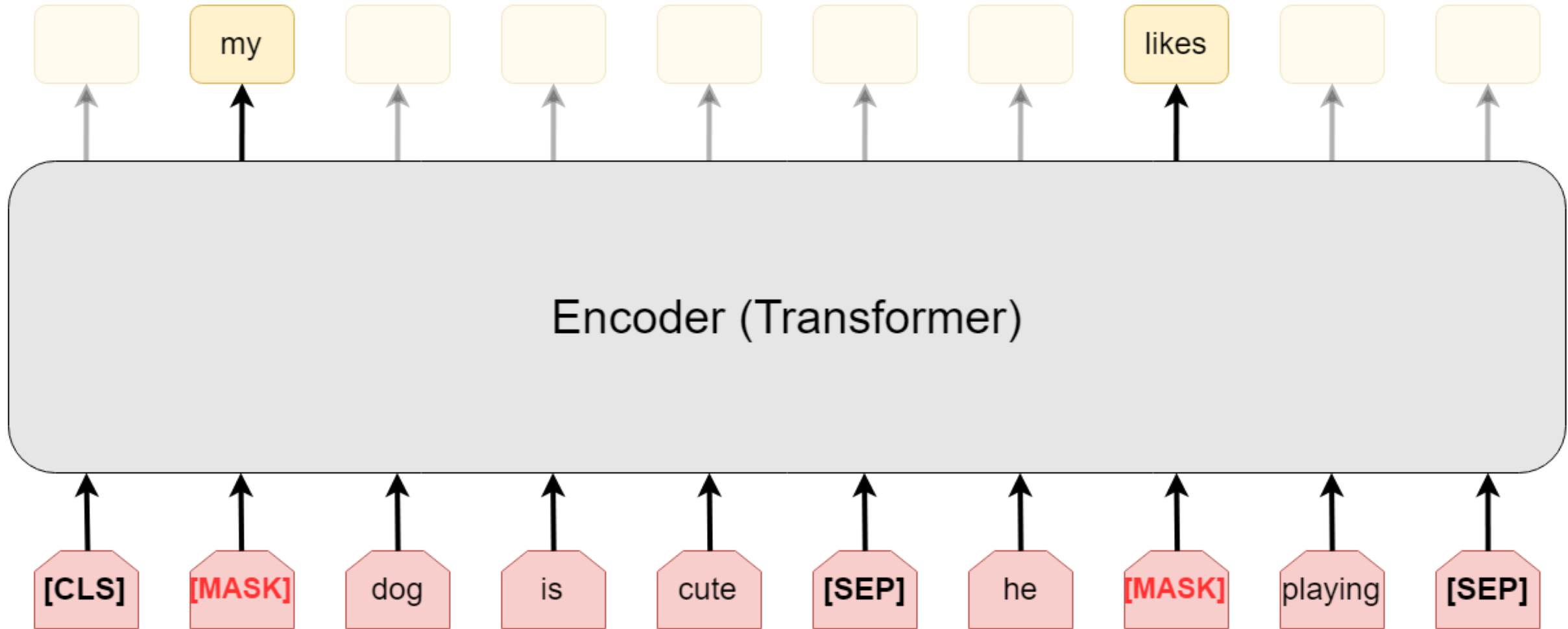


Feeding our static representation (LS) into a biLSTM-CRF model was SOTA in 2017 ([Ontonotes](#): 87.95 [CONLL](#): 91.73 F1-scores)

Contextual embeddings: ELMo (Peters et al. 2018)



Contextual embeddings: BERT (Devlin et al., 2018)

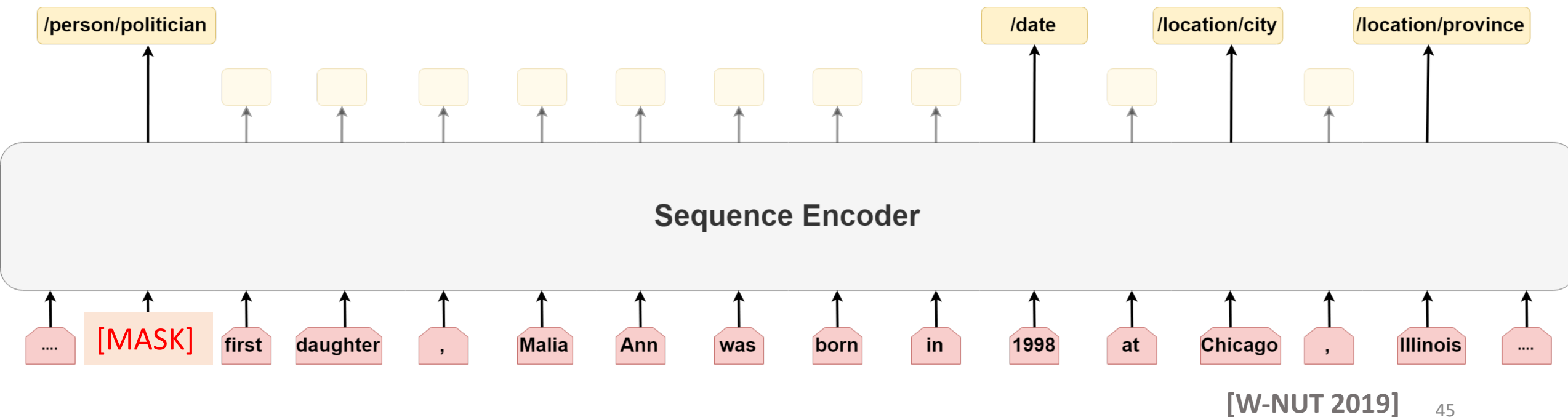


GhAWi: our Contextual Representation

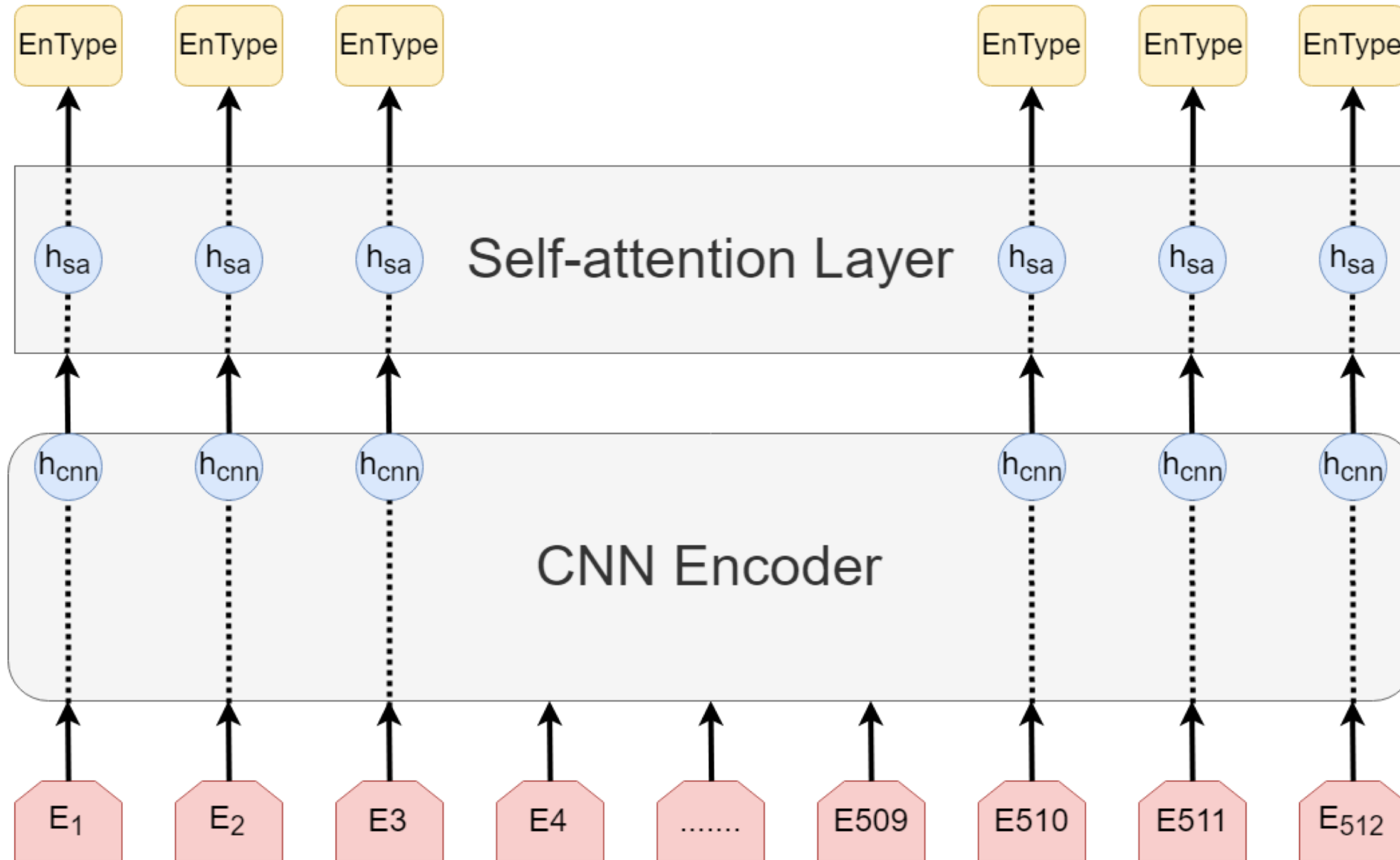
BERT inspired !

Sequences of 512 words max. (paragraph/section)

We mask some entities



GhAWi: a two-Component model

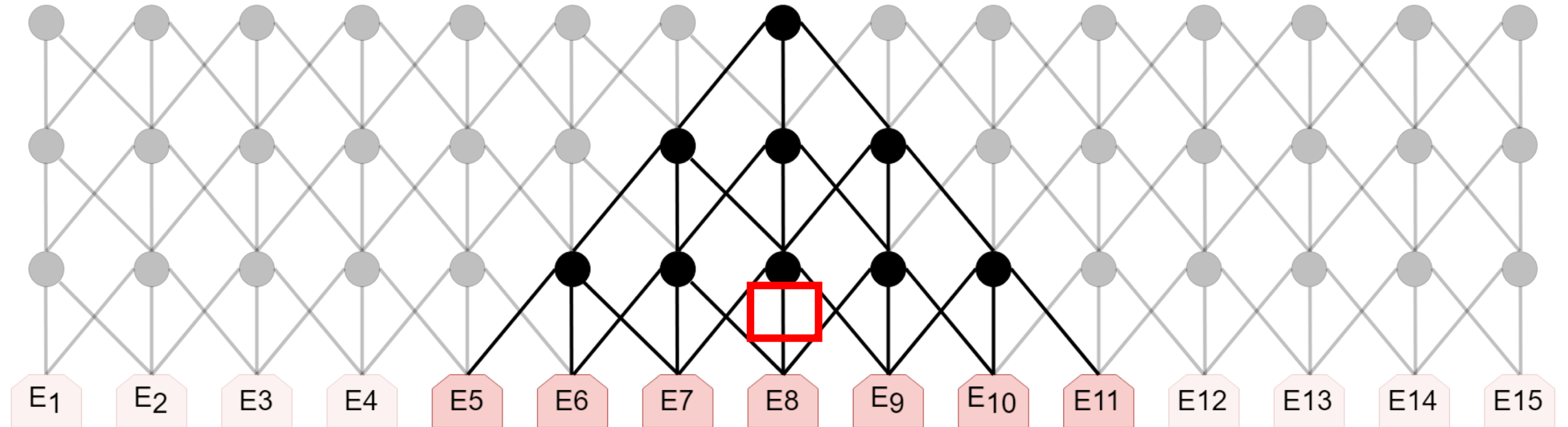


5 days of
training on a
Titan XP
GPU

Encoder: CNN

$$c_t = W_c \bigoplus_{k=0}^r x_{t \pm k}$$

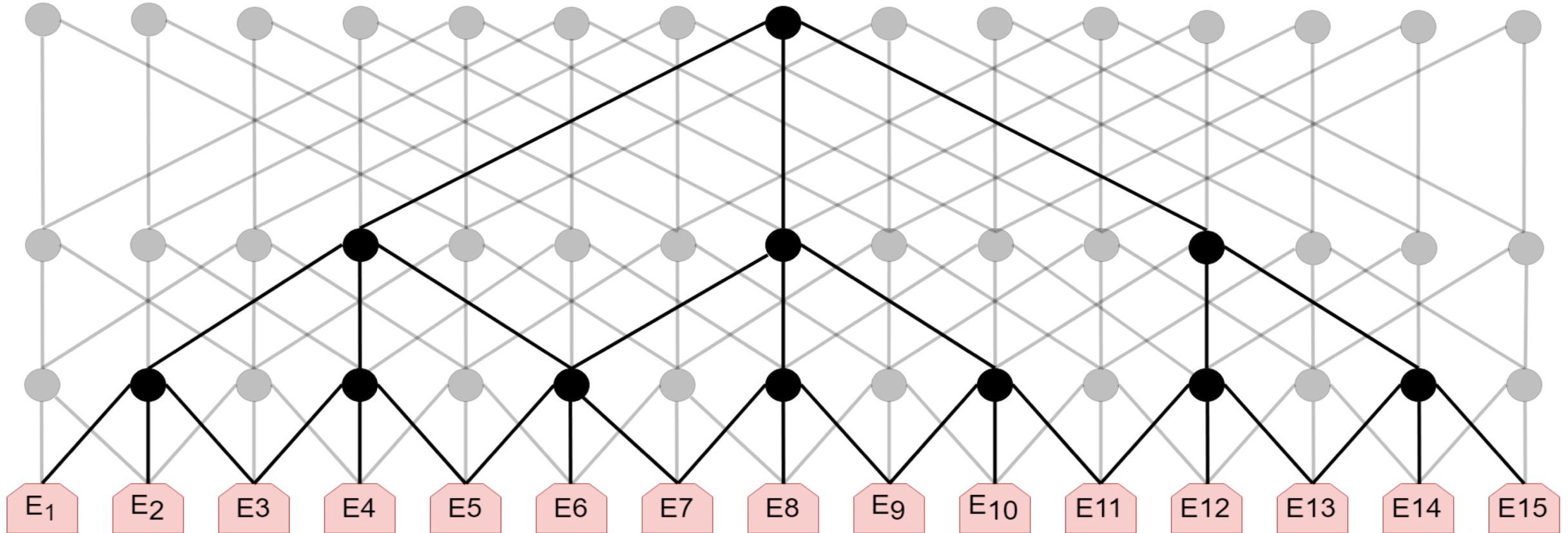
$$c_8 = W_c [E7; E8; E9]$$



255 layers to cover a batch of **512 words** with $r=3$

Dilated CNN (with increasing window size)

Yu and Koltum (2015), Strubbel et al. (2017)



« Only » need **8 layers** to cover **512 words**

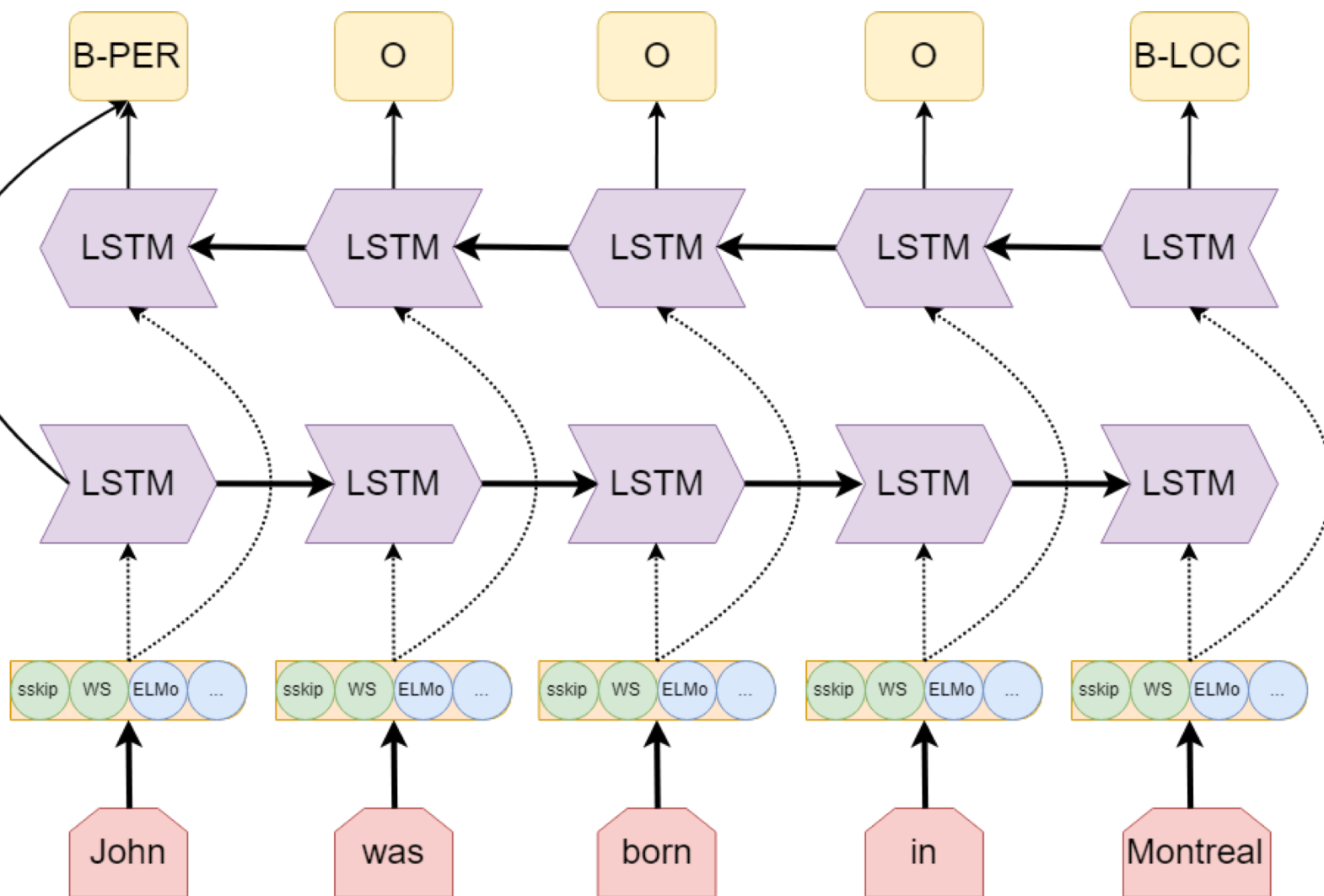
Encoder: Self-attention (Transformer)

| | John | was | born | in | Mtl |
|----------|------|------|------|------|------|
| John | 0.06 | 0.12 | 0.76 | 0.04 | 0.02 |
| was | * | * | * | * | * |
| born | * | * | * | * | * |
| in | * | * | * | * | * |
| Montreal | 0.08 | 0.09 | 0.37 | 0.41 | 0.05 |

born strongly suggests John is a person

born in strongly suggests Montreal is a location

Our NER recognizer



Baseline:

- sskip: skipgram model
- WS: Word Shape features

Contextual representations:

- ELMo
- Flair (Akbik et al. 2018)
- BERT
- GhAWi

Benchmarks

Onto



82k

18

CoNLL



24k

4

I2B2



12k

23

WNUT



2k

6

Fin



.5k

4

Wiki



H.D.

4

WP



H.D.

4

Adding contextual representations greedily

| | In-domain | | | | | Out-domain | |
|-----------------|-----------|-------|-------|-------|-------|------------|-------|
| | Onto | CoNLL | I2B2 | WNUT | Fin | Wiki | WP |
| SSKIP+WS | 86.44 | 90.73 | 86.41 | 32.30 | 81.82 | 66.03 | 45.13 |
| +ELMo | 89.37 | 92.47 | 94.47 | 44.15 | 82.03 | 76.34 | 54.45 |
| +GhAWi | 89.68 | 92.96 | 94.75 | 47.40 | 83.00 | 78.51 | 57.23 |
| +Flair | 89.73 | 93.22 | 94.79 | 46.80 | 83.11 | 77.77 | 56.20 |
| +Bert | 89.97 | 93.02 | 94.92 | 46.47 | 81.94 | 78.06 | 56.84 |

Comparing to SOTA

| | CoNLL | ONTO |
|-------------------------------|--------------|--------------|
| (Ghaddar et al., 2018) | 91.73 | 87.95 |
| (Peters et al., 2018) | 92.20 | - |
| (Clark et al., 2018) | 92.61 | 88.81 |
| (Devlin et al., 2018) | 92.80 | - |
| (Ghaddar et al., 2018) (best) | 92.87 | 89.69 |
| This work (best) | 93.22 | 89.95 |

Follow up

- Two very recent papers have been reproducing a similar approach
 - [Abhishek et al. 2019] Fine-grained Entity Recognition with Reduced False Negatives and Large Typw Coverage
 - [Zhu et al. 2019] Towards Open-Domain Named Entity Recognition via Neural Correction Models
- No direct comparison (yet), but clearly distant supervision is back again

Agenda

- Motivation: Open Information Extraction
- How robust is NER today?
- Better NER ?
 - Learning dedicated representations with distant supervision
 - **Better sequence labelling with multi-tasking**
- Conclusion

SC-LSTM: Learning Task-Specific Representations In Multi-Task Learning For Sequence Labeling

[NAACL 2019]



Peng Lu



Ting Bai

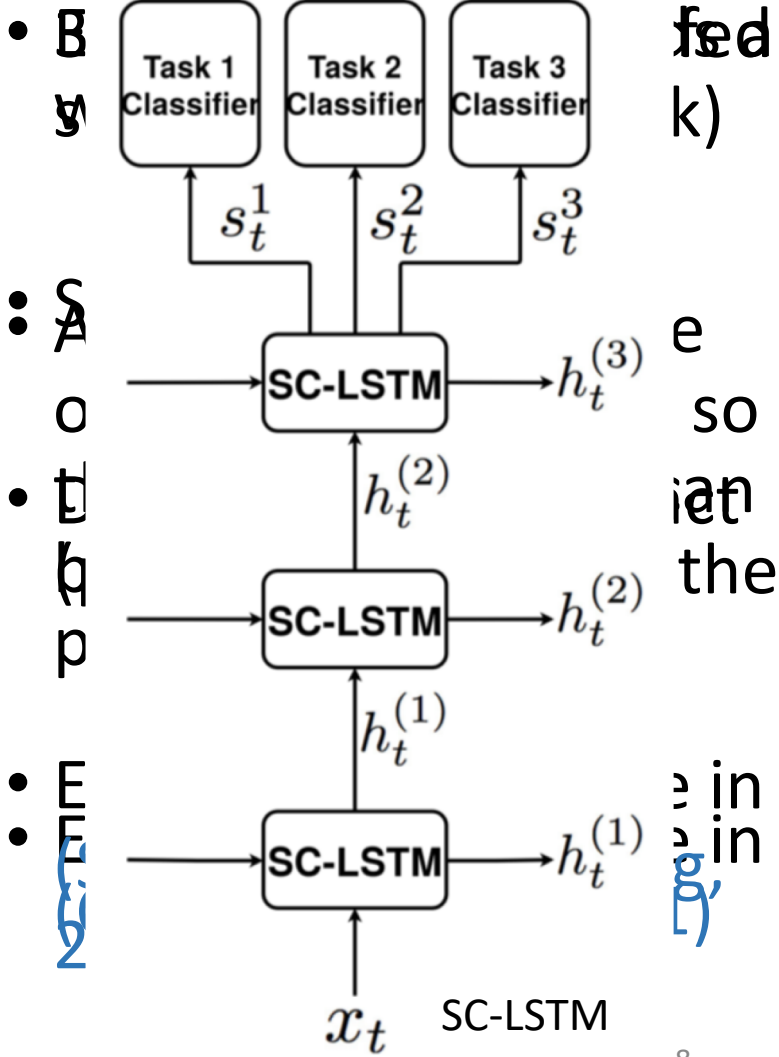
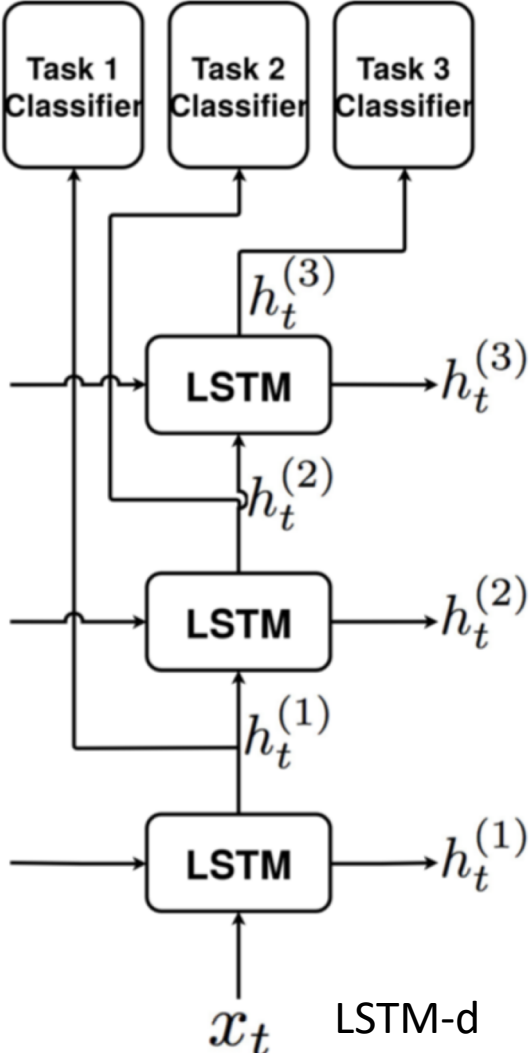
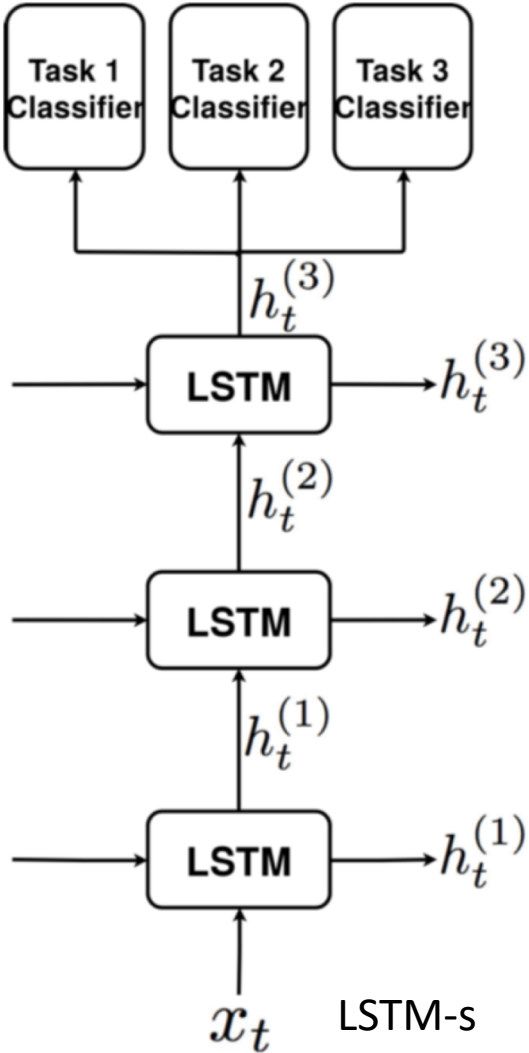
Jointly learning multiple tasks (MTL)

- Many attempts ([Caruana, 1997](#); [Collobert and Weston, 2008](#); [Collobert et al. 2011](#))
- Results so far are inconclusive
 - Works for some tasks, not for others ([ex: Alonso et Plank, 2017](#))
 - Sensitive to critical choices ([ex: order on the task being learned](#))
 - Often, one system has to be trained for a specific target task

Intuition: different tasks may have conflicting needs at training

=> Allowing task-specific parameters

MTL architectures



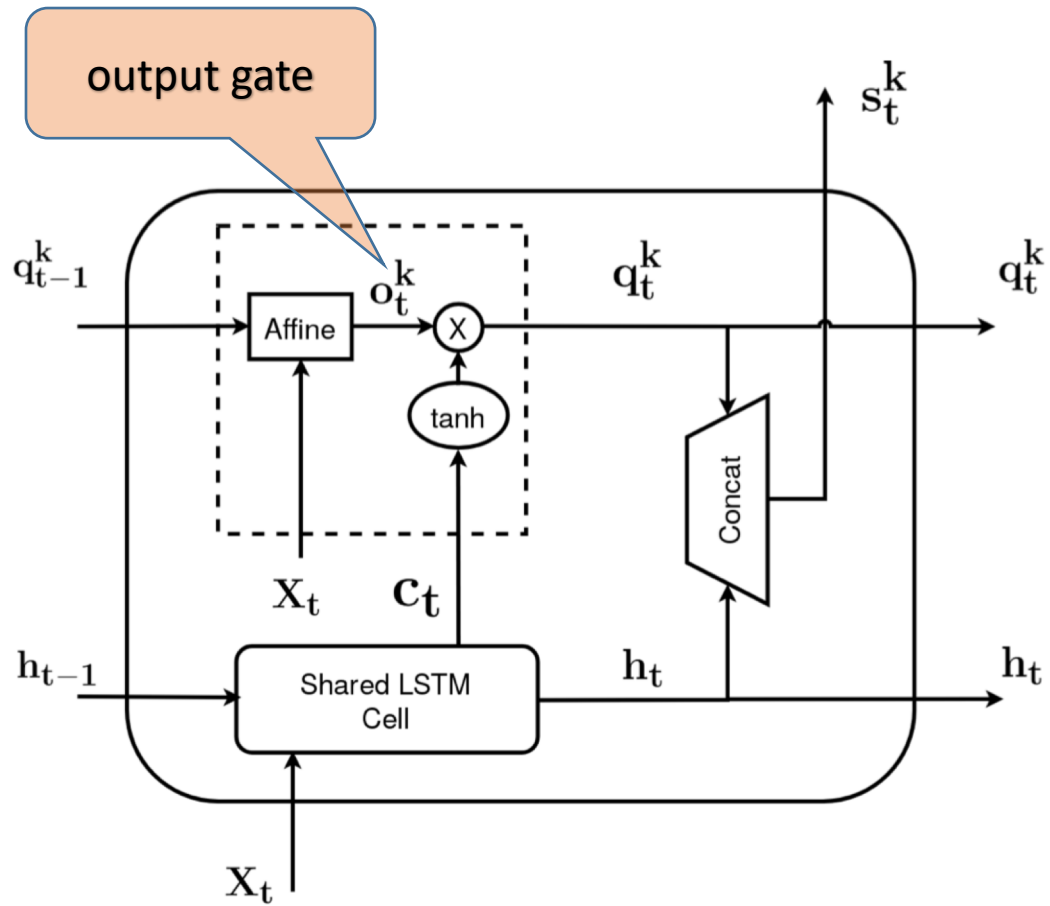
- By
- A
- o
- t
- h
- p
- E
- F

fed
k)

e
so
ian
the

in
in
by

SC-LSTM Cell



q_t^k = task specific representation

h_t = shared representation

task-specific matrices

$$\mathbf{o}_t^k = \sigma(\mathbf{W}_k \mathbf{q}_{t-1}^k + \mathbf{U}_k \mathbf{x}_t^k)$$

$$\mathbf{q}_t^k = \mathbf{o}_t^k * \tanh(\mathbf{c}_t)$$

Results

| | POS (accuracy) | Chunking (F1) | NER (F1) |
|-----------------------|----------------|---------------|--------------|
| LSTM (single task) | 95.46 | 94.44 | 89.39 |
| LSTM-s | 95.45 | 95.12 | 89.35 |
| LSTM-d | 95.44 | 95.24 | 89.37 |
| SC-LSTM | 95.51 | 96.04 | 89.96 |
| + char CNN & CRF | 95.83 | 96.41 | 91.37 |
| + char CNN & LM | 96.83 | 97.40 | 92.60 |
| (Devlin et al, 2018) | | | 92.80 |
| (Akbik et al., 2018) | | 96.72 | 93.09* |
| (Peters et al., 2018) | 96.62 | 96.92 | 92.22 |

Follow up

- At the very same conference
 - [Guo et al. 2019] AUTOSEM: Automatic Task Selection and Mixing in Multi-Task Learning
 - A two-stage approach
 - Selecting (automatically) the most usefull auxiliary tasks
 - Learning to mix them during training

Summary

- RALI is working on Open Information Extraction
 - Not easy to evaluate
 - Interplay with NER
- Most works on NER focuss on CoNLL and Ontonotes
 - Curiously not much interest in measuring NER in an out-domain setting (the one that matters)
- Distant supervision can help (to improve upon strong models)
- Multi-tasking may help as well

Thanks
Questions ?