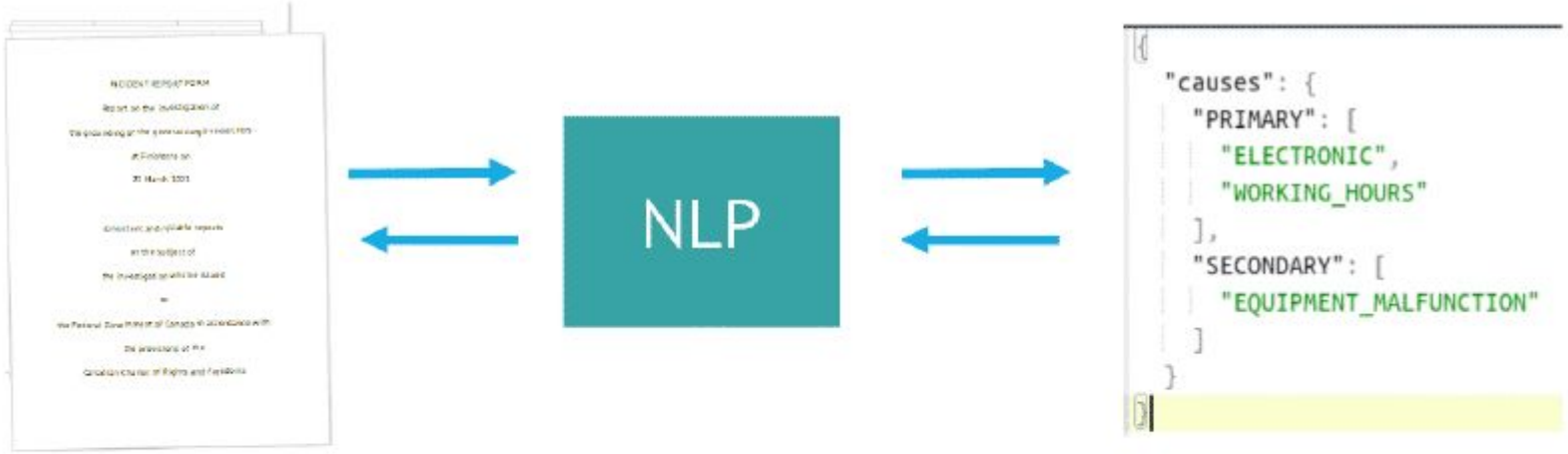
Natural Language Processing (Almost) from Scratch

Ronan Collobert[*] RONAN@COLLOBERT.COM
Jason Weston[†] JWESTON@GOOGLE.COM
Léon Bottou[‡] LEON@BOTTOU.ORG
Michael Karlen MICHAEL.KARLEN@GMAIL.COM
Koray Kavukcuoglu[§] KORAY@CS.NYU.EDU
Pavel Kuksa[¶] PKUKSA@CS.RUTGERS.EDU

Paper report by Nikita Ivlev, Lev Leontev, Alexander Kariakin

Introduction



Unstructured document

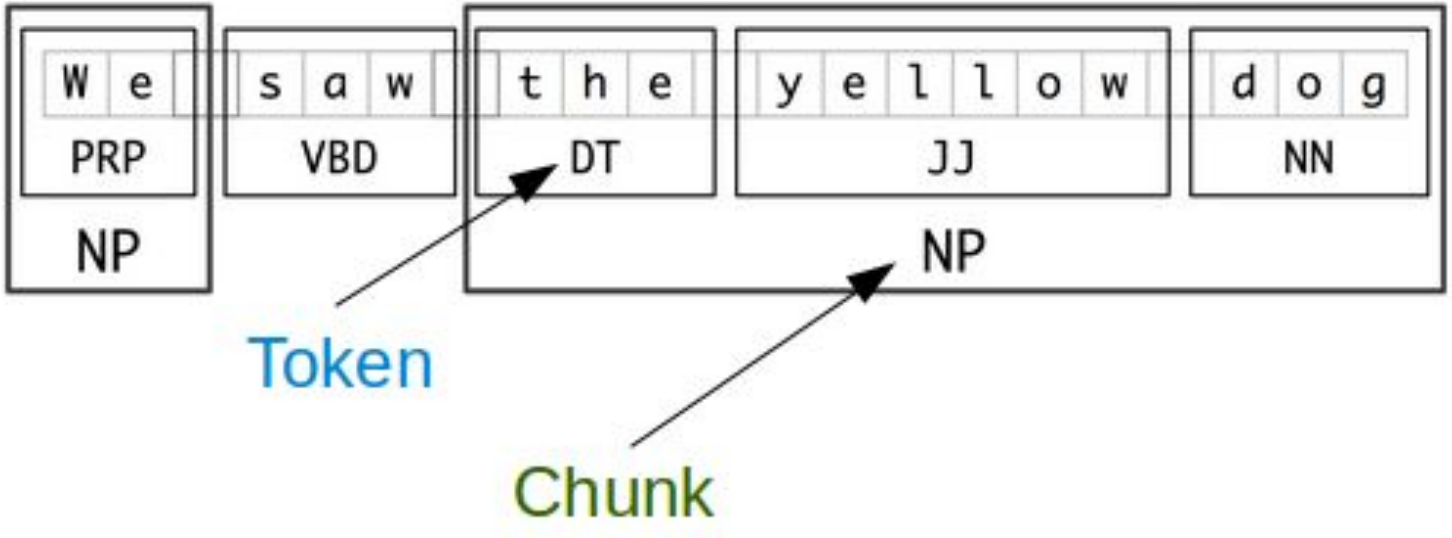
Structured format

Part-Of-Speech Tagging

Why	not	tell	someone	?
adverb	adverb	verb	noun	punctuation mark, sentence closer

Chunking

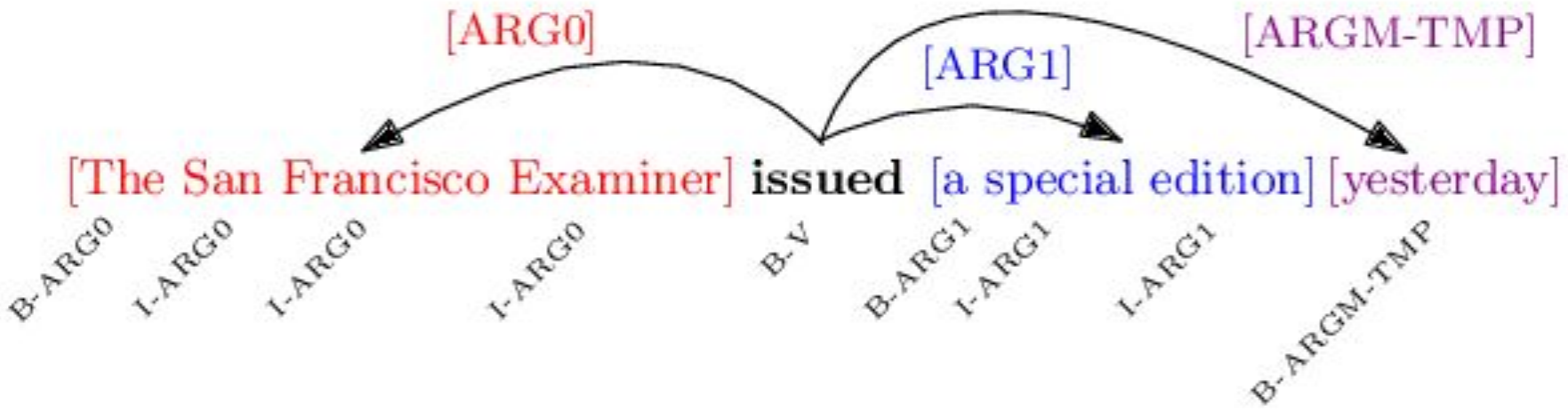
Also called shallow parsing, chunking aims at labeling segments of a sentence with syntactic constituents such as noun or verb phrases (NP or VP)



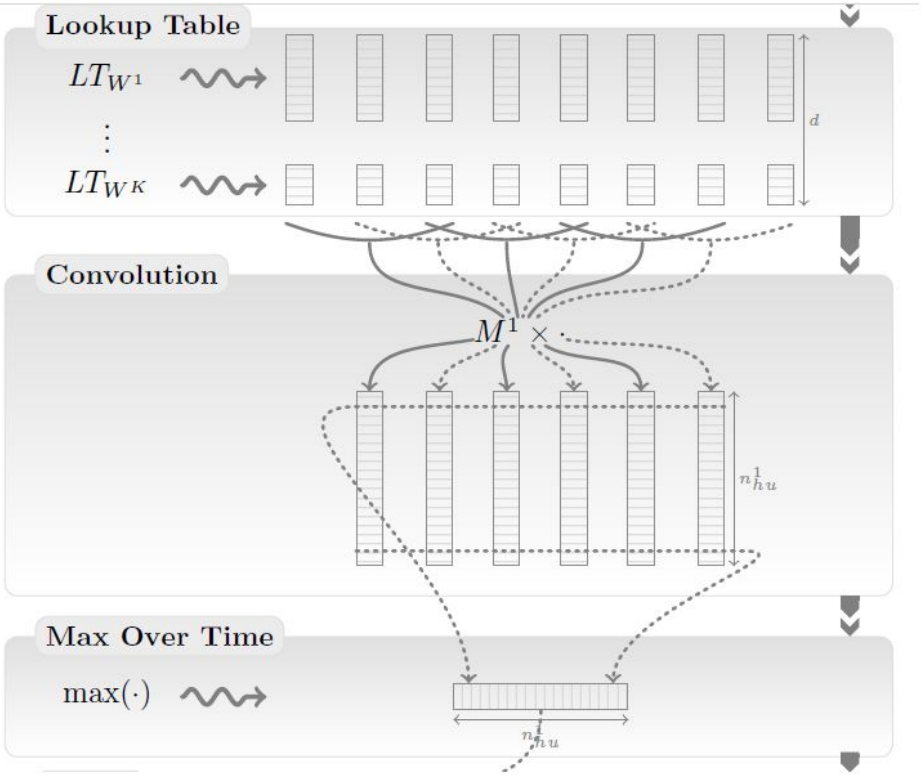
Named Entity Recognition

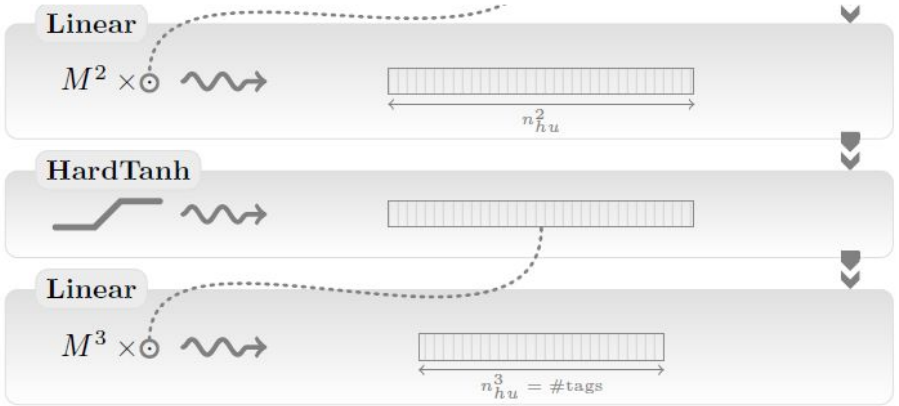
Semantic Role Labeling

SRL aims at giving a semantic role to a syntactic constituent of a sentence. For example, assigning roles ARG0-5 to words that are arguments of a verb (or a predicate) in the sentence.



Network architecture



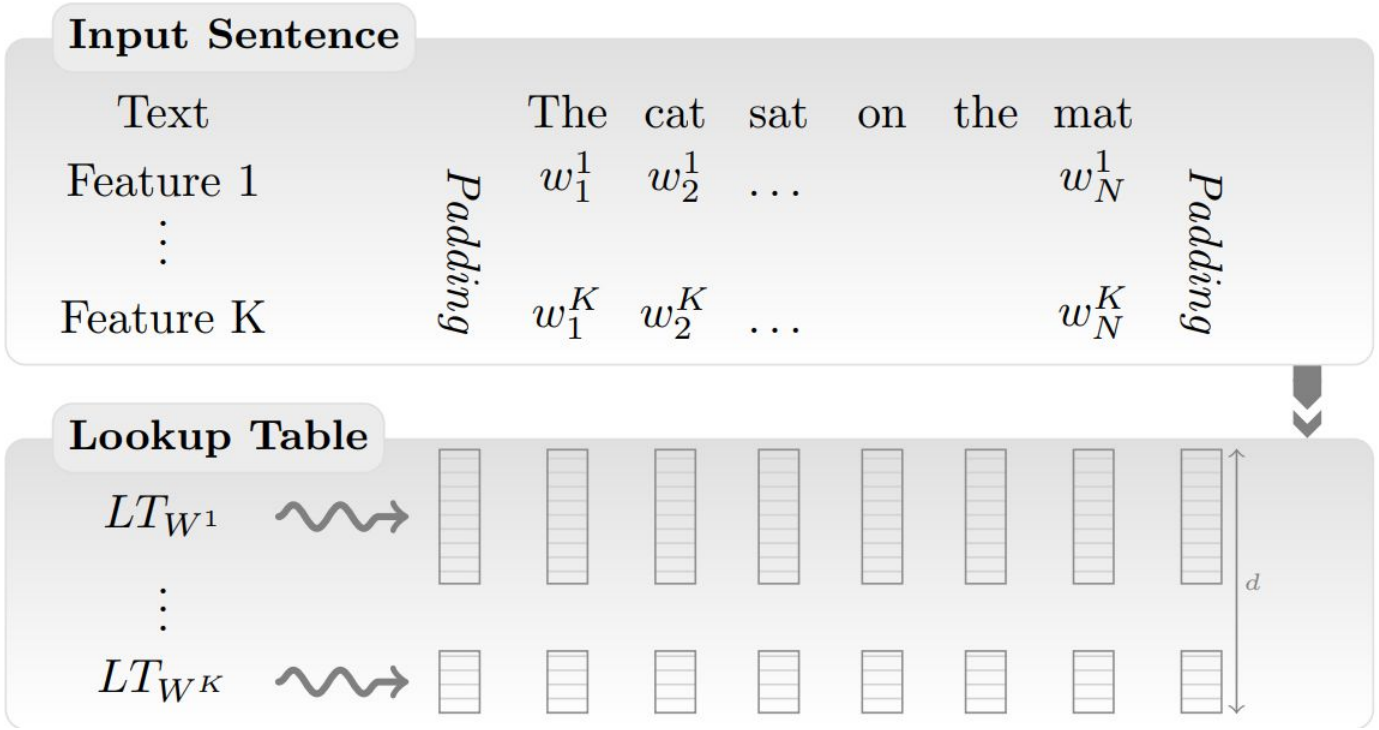


Transforming Words into Feature Vectors

$$D = \{\text{The}, \text{cat}, \text{sat}\}$$

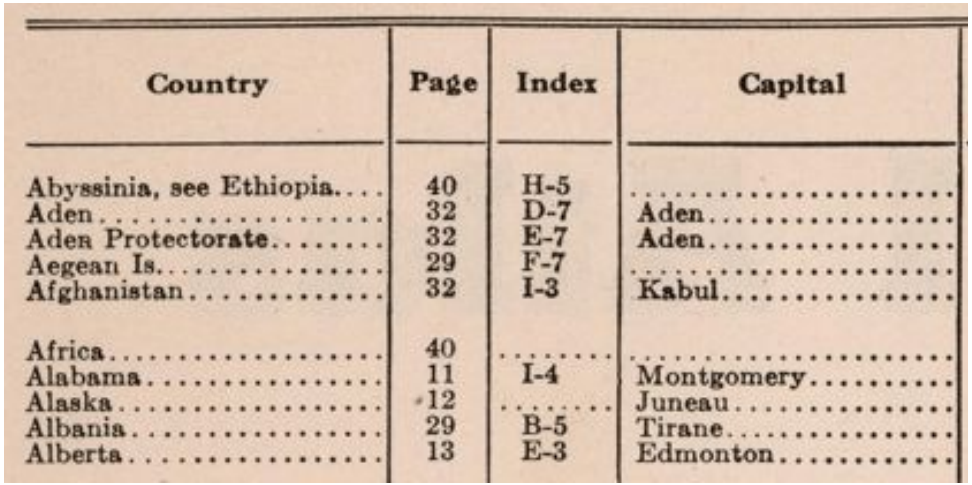
$$LT_w(\text{cat}) = \begin{pmatrix} 0.5 \\ 0.7 \\ 0.1 \end{pmatrix}$$

Lookup Table Layer

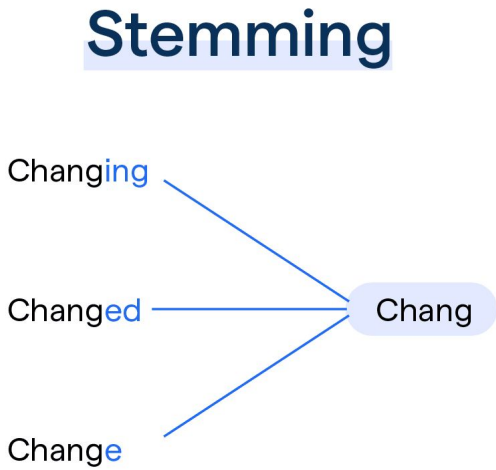


Extending to any discrete features

- Gazetteer
- Some basic pre-processing, such as word-stemming or dealing with upper and lower case



Country	Page	Index	Capital
Abyssinia, see Ethiopia	40	H-5	
Aden	32	D-7	Aden
Aden Protectorate	32	E-7	Aden
Aegean Is.	29	F-7	
Afghanistan	32	I-3	Kabul
Africa	40		
Alabama	11	I-4	Montgomery
Alaska	12		Juneau
Albania	29	B-5	Tirane
Alberta	13	E-3	Edmonton

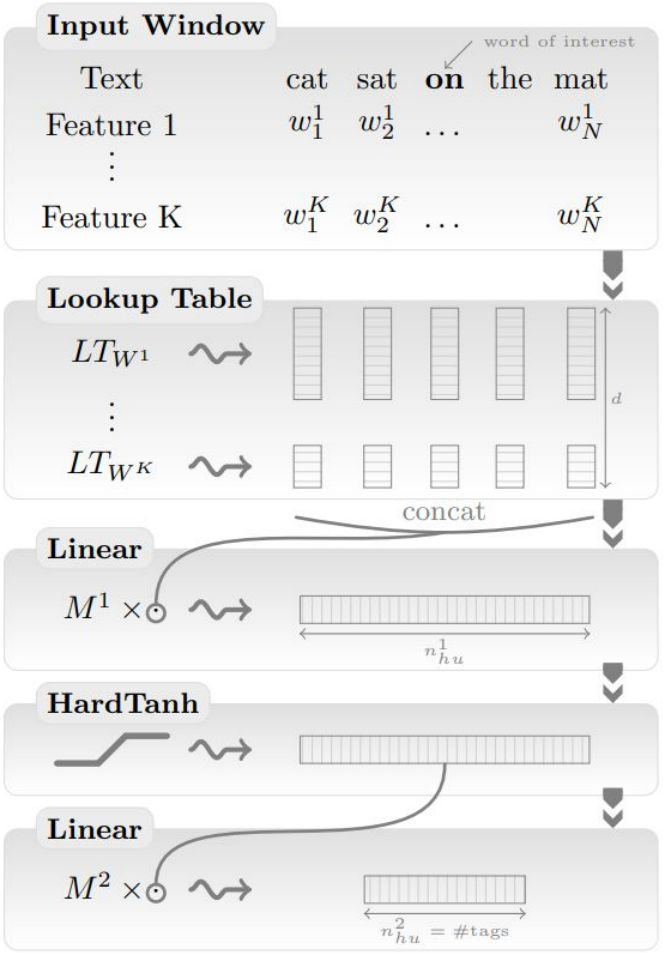


Window approach

Window approach network

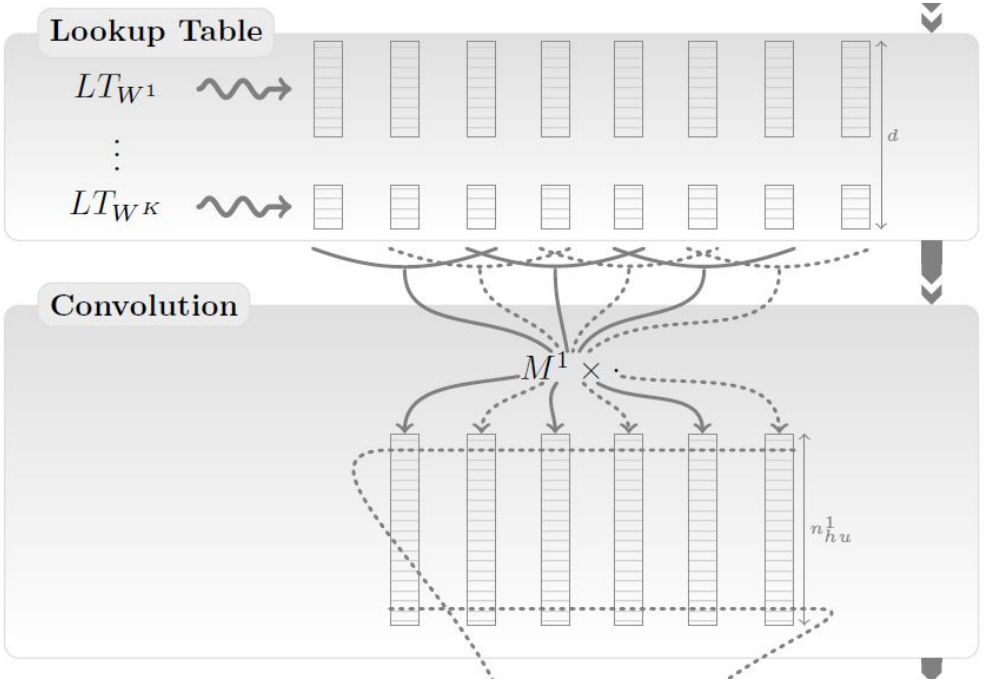
$$f_\theta^l = W^l f_\theta^{l-1} + b^l$$

$$\text{HardTanh}(x) = \begin{cases} -1 & \text{if } x < -1 \\ x & \text{if } -1 <= x <= 1 \\ 1 & \text{if } x > 1 \end{cases}$$



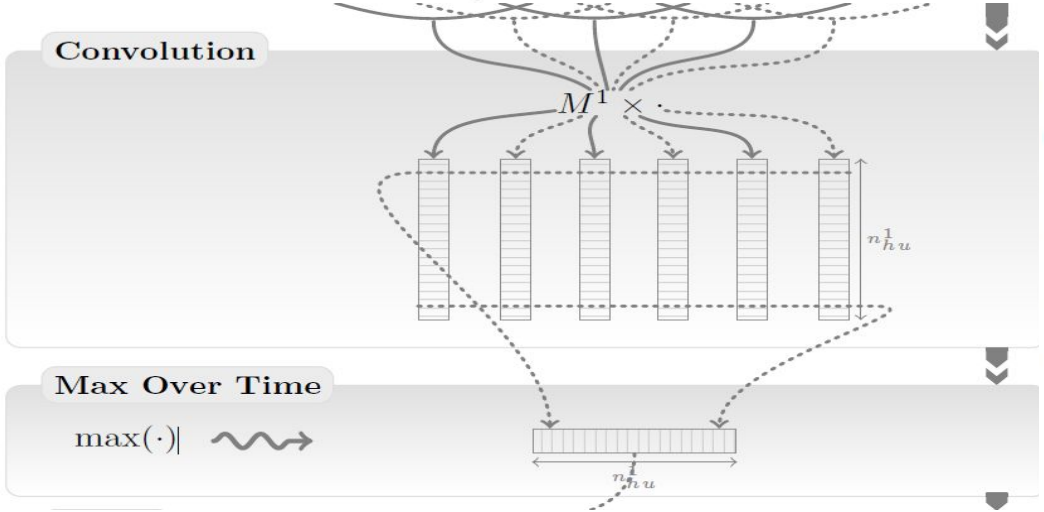
Convolutional Layer

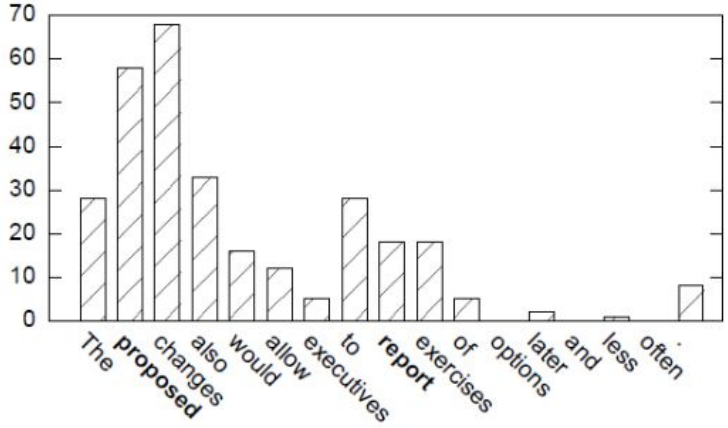
$$\langle f_\theta^l \rangle_t^1 = W^l \langle f_\theta^{l-1} \rangle_t^{d_{win}} + b^l$$

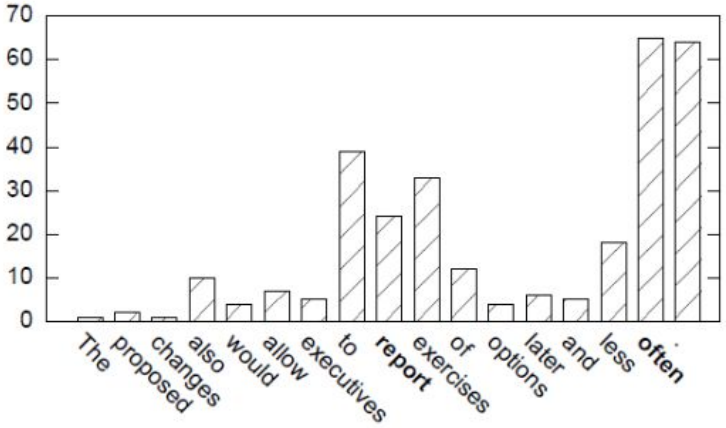


Max Layer

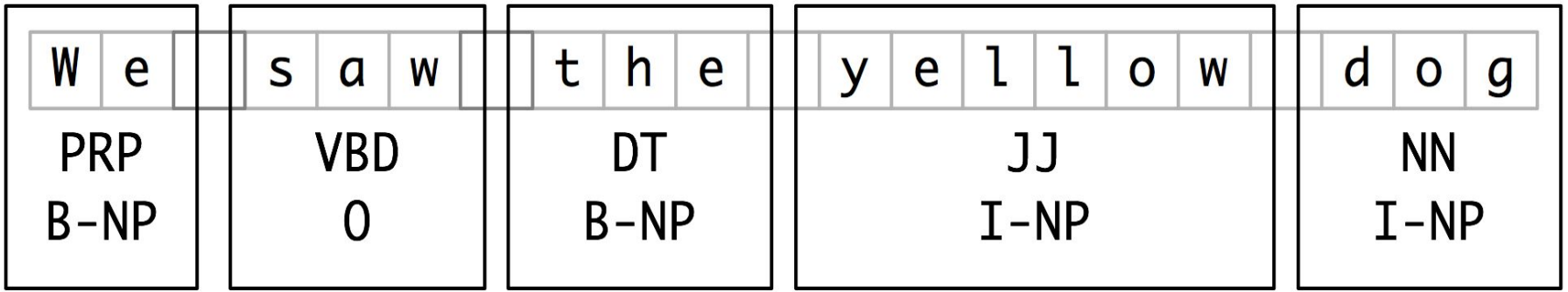
$$\left[f_\theta^l\right]_i = \max_t \left[f_\theta^{l-1}\right]_{i,t} \qquad 1 \le i \le n_{hu}^{l-1} .$$







Tagging schemes



Results

Approach	POS (PWA)	CHUNK (F1)	NER (F1)	SRL (F1)
Benchmark Systems	97.24	94.29	89.31	77.92
Window Approach				
NN+SLL+LM2	97.20	93.63	88.67	–
NN+SLL+LM2+MTL	97.22	94.10	88.62	–
Sentence Approach				
NN+SLL+LM2	97.12	93.37	88.78	74.15
NN+SLL+LM2+MTL	97.22	93.75	88.27	74.29

Approach	POS (PWA)	CHUNK (F1)	NER (F1)	SRL
Benchmark Systems	97.24	94.29	89.31	77.92
NN+SLL+LM2	97.20	93.63	88.67	74.15
NN+SLL+LM2+Suffix2	97.29	–	–	–
NN+SLL+LM2+Gazetteer	–	–	89.59	–
NN+SLL+LM2+POS	–	94.32	88.67	–
NN+SLL+LM2+CHUNK	–	–	–	74.72

Instead of exploiting man-made input features carefully optimized for each task, the system learns internal representations on the basis of vast amounts of mostly unlabeled training data. This work is then used as a basis for building a freely available tagging system with good performance and minimal computational requirements.