



Information Extraction

Sourangshu Bhattacharya



Problems

- Entity extraction

- Named entity recognition.

- Entity Linking or Entity Disambiguation.

- Finetype classification or Typing.

- Relation extraction

Named Entity Recognition

Germany's representative to the European Union's veterinary committee Werner Zwingman said on Wednesday consumers should ...

IL-2 gene expression and NF-kappa B activation through CD28 requires reactive oxygen production by 5-lipoxygenase.

Why NER?

- Question Answering
- Textual Entailment
- Co-reference Resolution
- Computational Semantics
- ...

NER Data/Bake-Offs

- CoNLL-2002 and CoNLL-2003 (British newswire)

- Multiple languages: Spanish, Dutch, English, German

- 4 entities: Person, Location, Organization, Misc

- MUC-6 and MUC-7 (American newswire)

- 7 entities: Person, Location, Organization, Time, Date, Percent, Money

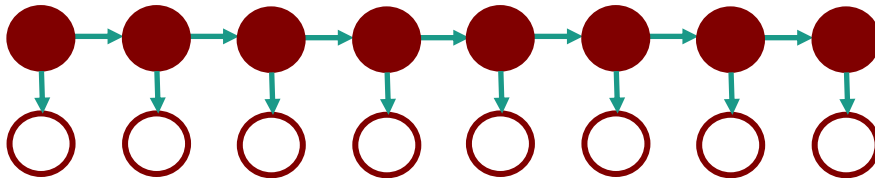
- ACE

- 5 entities: Location, Organization, Person, FAC, GPE

- BBN (Penn Treebank)

- 22 entities: Animal, Cardinal, Date, Disease, ...

Hidden Markov Models (HMMs)



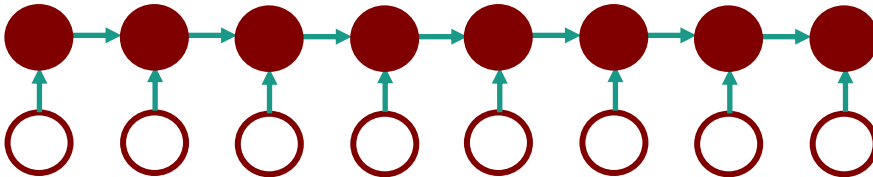
- Generative

- Find parameters to maximize $P(X,Y)$

- Assumes features are independent

- When labeling X_i future observations are taken into account (forward-backward)

MaxEnt Markov Models (MEMMs)



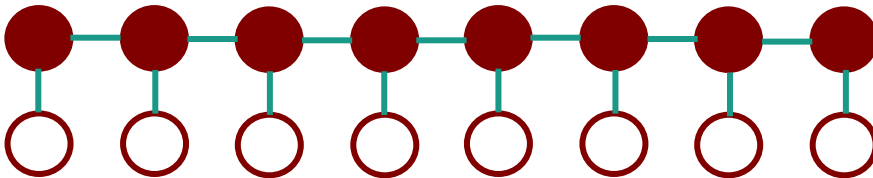
- Discriminative

○ Find parameters to maximize $P(Y|X)$

- No longer assume that features are independent

- Do not take future observations into account (no forward-backward)

Conditional Random Fields (CRFs)



- Discriminative
 - Doesn't assume that features are independent
 - When labeling Y_i future observations are taken into account
- The best of both worlds!

Model Trade-offs

	Speed	Discrim vs. Generative	Normalization
HMM	very fast	generative	local
MEMM	mid-range	discriminative	local
CRF	kinda slow	discriminative	global

NER-CRF features



- Whether to word starts with an uppercase letter
- Character length of a word
- Whether the word contains any digit (0-9)
- Whether the word contains any punctuation, i.e. . , ; () [] ? !
- Whether the word contains *only* digits
- Whether the word contains *only* punctuation
- The [word2vec](#) cluster of the word (add -classes flag to the word2vec tool)
- ...

Stanford NER features



Feature	NER	TF
Current Word	✓	✓
Previous Word	✓	✓
Next Word	✓	✓
Current Word Character n-gram	all	length ≤ 6
Current POS Tag	✓	
Surrounding POS Tag Sequence	✓	
Current Word Shape	✓	✓
Surrounding Word Shape Sequence	✓	✓
Presence of Word in Left Window	size 4	size 9
Presence of Word in Right Window	size 4	size 9

Our Features

- Word features: current word, previous word, next word, all words within a window

- Orthographic features:

oJenny Xxxx

oIL-2 XX-#



- Prefixes and Suffixes:

oJenny <J, <Je, <Jen, ..., nny>, ny>, y>



- Label sequences

- Lots of feature conjunctions

Distributional Similarity Features

- Large, unannotated corpus
- Each word will appear in contexts - induce a distribution over contexts
- Cluster words based on how similar their distributions are
- Use cluster IDs as features
- Great way to combat sparsity
- We used Alexander Clark's distributional similarity code (easy to use, works great!)
- 200 clusters, used 100 million words from English gigaword corpus

Deep learning based model (EMNLP 16)



NeuroNER: an easy-to-use program for named-entity recognition based on neural networks

Franck Deroncourt*
MIT
francky@mit.edu

Ji Young Lee*
MIT
jjylee@mit.edu

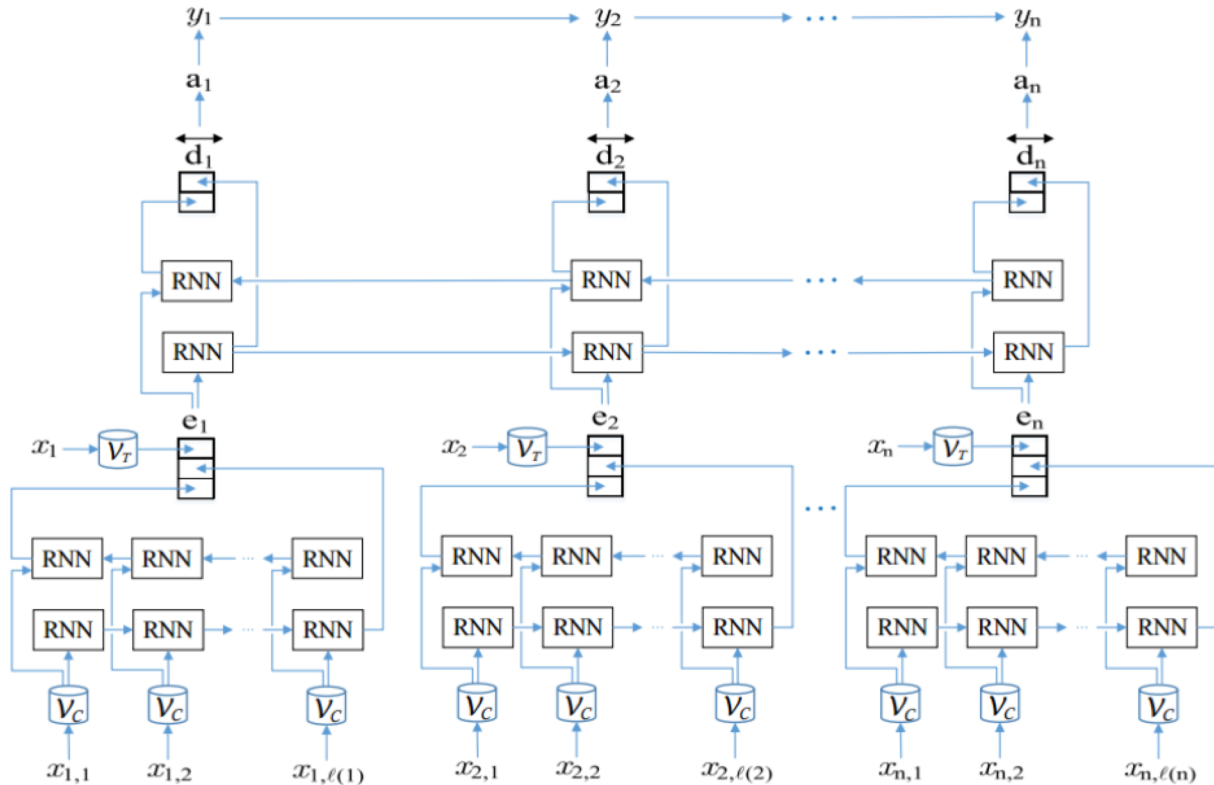
Peter Szolovits
MIT
psz@mit.edu

Abstract

Named-entity recognition (NER) aims at identifying entities of interest in a text. Artificial neural networks (ANNs) have recently been shown to outperform existing NER systems. However, ANNs remain challenging to use for non-expert users. In this paper, we present NeuroNER, an easy-to-use named-entity recognition tool based on ANNs. Users can annotate entities using a graphical web-based user interface

Fully supervised approaches to NER include support vector machines (SVM) ([Asahara and Matsumoto, 2003](#)), maximum entropy models ([Borthwick et al., 1998](#)), decision trees ([Sekine et al., 1998](#)) as well as sequential tagging methods such as hidden Markov models ([Bikel et al., 1997](#)), Markov maximum entropy models ([Kumar and Bhattacharyya, 2006](#)), and conditional random fields (CRFs) ([McCallum and Li, 2003](#); [Tsai et al., 2006](#); [Benajiba and Rosso, 2008](#); [Filannino et al., 2013](#)). Similar to rule-based systems, these

Neuro_NER



Neuro_NER



Figure 1: Architecture of the artificial neural network (ANN) model. RNN stands for recurrent neural network. The type of RNN used in this model is Long Short Term Memory (LSTM). n is the number of tokens, and x_i is the i^{th} token. \mathcal{V}_T is the mapping from tokens to token embeddings. $\ell(i)$ is the number of characters and $x_{i,j}$ is the j^{th} character in the i^{th} token. \mathcal{V}_C is the mapping from characters to character embeddings. \mathbf{e}_i is the character-enhanced token embeddings of the i^{th} token. $\overleftrightarrow{\mathbf{d}}_i$ is the output of the LSTM of label prediction layer, \mathbf{a}_i is the probability vector over labels, y_i is the predicted label of the i^{th} token.




Named Entity Disambiguation

Introduction

- Named Entity Disambiguation is a central problem of Information Extraction where the goal is to link entities in a knowledge base (KB) to their mention spans in unstructured text.
- A knowledge base (KB) is a technology used to store complex structured and unstructured information used by a computer system.

Michael
Jordan(Basketball
Player)

Michael
Jordan(Statistician)



Michael Jordan is an American retired professional basketball player, businessman, and principal owner and chairman of the Charlotte Hornets. Jordan played 15 seasons in the National Basketball Association (NBA) for the Chicago Bulls and Washington Wizards.

Robust Disambiguation of Named Entities in Text

[EMNLP'11]



- Graph-based Approach to solve NED.
- Consider an input text (Web page, news article, blog posting, etc.) and find out noun phrases using NER Tagger.
- Collect candidate entities and their prior for each mention.
- Measure Context Similarity between entity context and mention context.
- Measure Entity-Entity Coherence by the number of incoming links that their Wikipedia articles share.

Robust Disambiguation of Named Entities in Text

[EMNLP'11]



- We can create **mention-entity graph** where
 - Nodes are mentions and entities.
 - Mention-Entity edge is weighted with a combination of prior and similarity measure between mention and entity
 - Entity-entity edge is weighted based on Entity-Entity coherence.

$$\alpha \cdot \sum_{i=1..k} \text{prior}(m_i, e_{j_i}) +$$
$$\beta \cdot \sum_{i=1..k} \text{sim}(\text{cxt}(m_i), \text{cxt}(e_{j_i})) +$$
$$\gamma \cdot \text{coh}(e_{j_1} \in \text{cnd}(m_1) \dots e_{j_k} \in \text{cnd}(m_k))$$

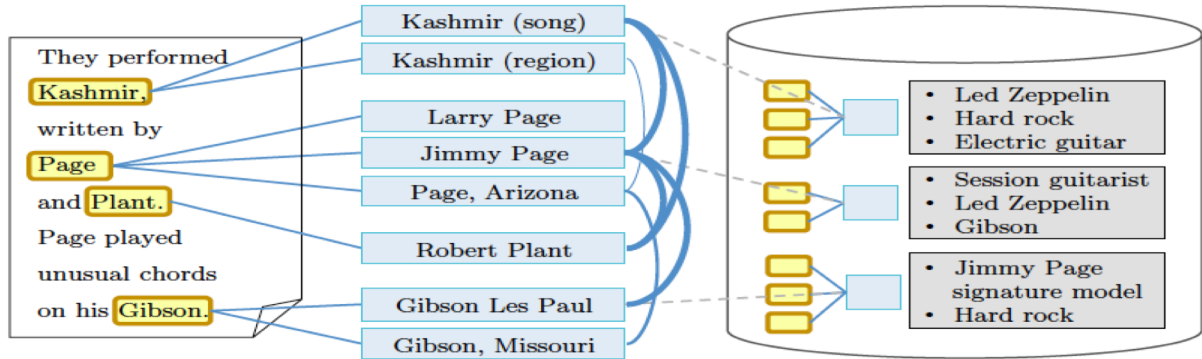
Robust Disambiguation of Named Entities in Text

[EMNLP'11]



- The density of a subgraph to be equal to the minimum weighted degree among its nodes.
- Given a mention-entity graph, goal is to compute a dense subgraph that would ideally contain all mention nodes and exactly one mention-entity edge for each mention, thus disambiguating all mentions.

Robust Disambiguation of Named Entities in Text [EMNLP'11]



Mention-Entity Graph

Robust Disambiguation of Named Entities in Text [EMNLP'11]



An entity is taboo if it is the last candidate for a mention it is connected to.

Graph Disambiguation Algorithm:

Input: Mention-Entity Weighted Graph

Output : result graph with one edge per mention

1. foreach entity do
 - a. Calculate distance to all mentions.
2. keep the closest ($5 * \text{mentions count}$) entities, drop the others;
3. while graph has non-taboo entity
 - a. determine non-taboo entity node with lowest weighted degree, remove it.
 - b. if minimum weighted degree increased
 - i. Set solution to current graph.
4. process solution by local search or full enumeration for best configuration

Robust Disambiguation of Named Entities in Text

[EMNLP'11]



- The algorithm starts from the full mention-entity graph and iteratively removes the entity node with the smallest weighted degree.
- Among the subgraphs obtained in the various steps, the one maximizing the minimum weighted degree will be returned as output.
- To guarantee that we arrive at a coherent mention-entity mapping for all mentions, we enforce each mention node to remain connected to at least one entity

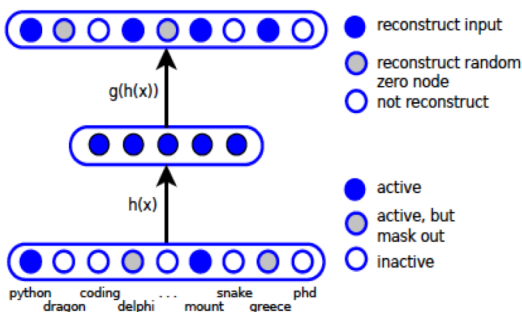
Learning Entity Representation for Entity Disambiguation [ACL'13]



- Deep Neural Network based approach to solve NED.
- Given a mention string m with its context document d , a list of candidate entities $C(m)$ are generated from, for each candidate entity $e_i \in C(m)$, we compute a ranking score $\text{sim}(d_m, e_i)$ indicating how likely m refers to e_i . The linking result is $e = \arg \max_{e_i} \text{sim}(d_m, e_i)$

Learning Entity Representation for Entity Disambiguation [ACL'13]

- The approach consists of two steps:
- Step 1: Greedy Layer-wise Pre-training :
 - Goal is to minimize reconstruction error $L(x, g(h(x)))$
 - Thus using Denoising Autoencoder to retain important information while ignoring noise.



Learning Entity Representation for Entity Disambiguation [ACL'13]



- Supervised Fine Tuning :

1. Similarity score of (d,e) pair is

- a. $L(d, e) = \max\{0, 1 - \text{sim}(d, e) + \text{sim}(d, e')\}$

2. Goal is to rank the correct entity higher than the rest candidates relative to the context of the mention.

$$L(d, e) = -\log \frac{\exp \text{sim}(d, e)}{\sum_{e_i \in C_m} \exp \text{sim}(d, e_i)}$$

Learning Entity Representation for Entity Disambiguation [ACL'13]



- Supervised Fine Tuning :
 3. For each training instance (d, e) , contrast it with one of its negative candidate pair (d, e') . This gives the pairwise ranking criterion
 - a. $\text{sim}(d,e)=\text{dot}(f(d),f(e))$
 4. Alternatively, we can contrast with all its candidate pairs (d, e_i) . That is, we raise the similarity score of true pair $\text{sim}(d, e)$ and penalize all the rest $\text{sim}(d, e_i)$. Then the loss function becomes

Minimize the following training objective across all training instances:

$$L = \sum_{d,e} L(d,e)$$

Learning Entity Representation for Entity Disambiguation [ACL'13]

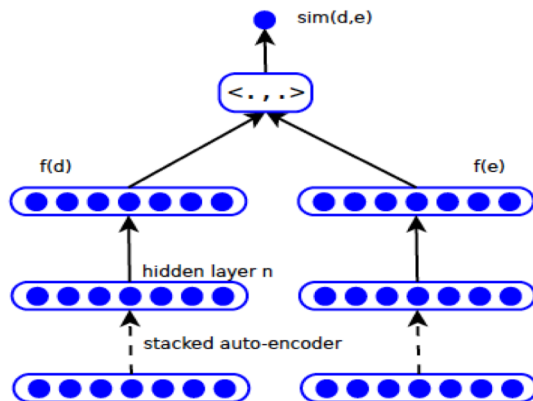


Figure 2: Network structure of fine-tuning stage.

Plato: A Selective Context Model for Entity Resolution[ACL'15]

- A probabilistic model for NED.
- Let us consider an example: *While **Solon** may have slipped slightly this year in **Cleveland** magazine's ranking of best suburbs, it climbed higher on a more prestigious list. On Monday, the city placed 23rd on Money magazine's annual list of best places to live in the United States.*
- There are five US locations named Solon in Wikipedia. In the above, **Solon** refers to a suburb of **Cleveland, Ohio**.
- But the only context feature that helps us discriminate between the different possible disambiguations is **Cleveland**.
- Selective context model aims to address this issue.

Plato: A Selective Context Model for Entity Resolution[ACL'15]

- A simplifying modeling assumption is that exactly one context feature is relevant for disambiguation, and the remaining features are drawn from a background distribution.
- We use w_m to represent the phrase of mention m . The context of mention m is represented either as a binary feature presence vector $\mathbf{b}_m \in \{0, 1\}^{|\mathcal{F}|}$, or as a feature count vector $\mathbf{c}_m \in \mathbb{N}^{|\mathcal{F}|}$.
- Let K be the random variable corresponding to the index of the relevant feature for a mention of entity e . The model can be written as

$$p(k, e, w, \mathbf{b}) = p(w)p(e|w)p(k|e) \prod_j p(b_j|k)$$

$$p(b_j|k) = \begin{cases} b_j & \text{if } k = j \\ \beta_j^{b_j} (1 - \beta_j)^{1-b_j} & \text{if } k \neq j. \end{cases}$$

- The β_j parameterizes the background probability of observing a feature that is not relevant.

Plato: A Selective Context Model for Entity Resolution[ACL'15]

- Given a test mention (w', b') , we can compute the entity posterior by marginalizing out the latent variable k .

$$\begin{aligned} p(e|b', w') &\propto p(e|w') \sum_k p(k|e) b'_k \prod_{j \neq k} p(b'_j | K \neq j) \\ &\propto p(e|w') \sum_k b'_k \frac{p(k|e)}{\beta_k}. \end{aligned} \quad (2)$$



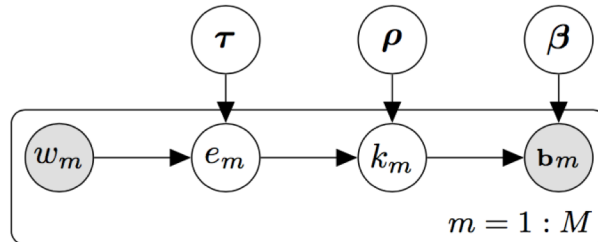
Plato

- Model

- vectors τ_w parameterize the conditional probability of an entity given the phrase w , with $\tau_{w,e} = p(E = e|W = w)$
- vectors ρ_e parameterize the probability of relevant features for entity e , with $\rho_{e,k} = p(K = k|E = e)$
- scalars $\beta_j = p(B_j = 1|K \neq j)$ parameterize the background feature distribution.

Plato

- Graphical model:



- Likelihood:

$$\begin{aligned} \mathcal{L} = & \sum_m \sum_e q_m(e) \left(\sum_w [w_m = w] \ln \tau_{e,w} \right. \\ & + \sum_k s_m(k) (\ln(b_{m,k} \rho_{k,e}) - \ln \beta_k) \\ & \left. + \sum_j b_{m,j} \ln \beta_j + (1 - b_{m,j}) \ln(1 - \beta_j) \right) \\ & + \mathcal{H}(q) + \mathcal{H}(s) + \text{const} \end{aligned}$$

Deep Joint Entity Disambiguation with Local Neural Attention[EMNLP'17]

- Deep Learning based approach to solve NED.
- **Learning Entity Embeddings:** Propose to train entity vectors that can be used for the ED task which compress semantic meaning of entities.
- The model embeds words and entities in the same low-dimensional vector space in order to exploit geometric similarity between them.
- Start with a pre-trained word embedding map $\mathbf{x} : W \rightarrow \mathbb{R}_d$ that is known to encode semantic meaning of words $w \in W$; specifically word2vec is used.
- Let $q(w)$ be a modified unigram word distribution $p'(w)$ which is for sampling "negative" words.
- Denote $w^+ \sim p'(w)$ and $w^- \sim q(w)$ and $q(w) = p'(w)^\alpha$

$$J(\mathbf{z}; e) = \mathbb{E}_{w^+|e} \mathbb{E}_{w^-} [h(\mathbf{z}; w^+, w^-)]$$

$$h(\mathbf{z}; w, v) = [\gamma - \langle \mathbf{z}, \mathbf{x}_w - \mathbf{x}_v \rangle]_+$$

$$\mathbf{x}_e = \arg \min_{\mathbf{z}: \|\mathbf{z}\|=1} J(\mathbf{z}; e)$$

Deep Joint Entity Disambiguation with Local Neural Attention[EMNLP'17]

- Local Model with Neural Attention:
- For a word compute score as

$$u(w) = \max_{e \in \Gamma(m)} \mathbf{x}_e^\top \mathbf{A} \mathbf{x}_w$$

- The score is high, if a word is strongly related to at least one candidate entity, then hard pruned $R \leq K$ words which received the low score.

$$\beta(w) = \begin{cases} \frac{\exp[u(w)]}{\sum_{v \in \bar{c}} \exp[u(v)]} & \text{if } w \in \bar{c} \\ 0 & \text{otherwise.} \end{cases}$$

$$\Psi(e, c) = \sum_{w \in \bar{c}} \beta(w) \mathbf{x}_e^\top \mathbf{B} \mathbf{x}_w$$

Adding mention-entity prior for final score

$$\Psi(e, m, c) = f(\Psi(e, c), \log \hat{p}(e|m))$$

Deep Joint Entity Disambiguation with Local Neural Attention[EMNLP'17]

- **Document-Level Deep Model:** Address global ED assuming document coherence among entities.

- The proposed model uses fully-connected pairwise conditional random field, defined as

$$g(\mathbf{e}, \mathbf{m}, \mathbf{c}) = \sum_{i=1}^n \Psi(e_i, m_i, c_i) + \sum_{i < j} \Phi(e_i, e_j)$$

$$\Phi(e, e') = \frac{2}{n-1} \mathbf{x}_e^\top \mathbf{C} \mathbf{x}_{e'}, \quad \text{where } \mathbf{C} \text{ is a diagonal matrix.}$$

- The pairwise factors are bilinear forms of the entity embeddings.

Deep Joint Entity Disambiguation with Local Neural Attention[EMNLP'17]

Mention	Gold entity	$\hat{p}(e m)$ of gold entity	Attended contextual words
Scotland	Scotland national rugby union team	0.034	England Rugby team squad Murrayfield Twickenham national play Cup Saturday World game George following Italy week Friday selection dropped row month
Wolverhampton	Wolverhampton Wanderers F.C.	0.103	matches League Oxford Hull league Charlton Oldham Cambridge Sunderland Blackburn Sheffield Southampton Huddersfield Leeds Middlesbrough Reading Coventry Darlington Bradford Birmingham Enfield Barnsley
Montreal	Montreal Canadiens	0.021	League team Hockey Toronto Ottawa games Anaheim Edmonton Rangers Philadelphia Caps Buffalo Pittsburgh Chicago Louis National home Friday York Dallas Washington Ice
Santander	Santander Group	0.192	Carlos Telmex Mexico Mexican group firm market week Ponce debt shares buying Televisa earlier pesos share stepped Friday analysts ended
World Cup	FIS Alpine Ski World Cup	0.063	Alpine ski national slalom World Skiing Whistler downhill Cup events race consecutive weekend Mountain Canadian racing

Table 7: Examples of context words selected by our local attention mechanism. Distinct words are sorted decreasingly by attention weights and only words with non-zero weights are shown.