diaNED: Time-Aware Named Entity Disambiguation for Diachronic Corpora

Prabal Agarwal[1], Jannik Strötgen[1,2], Luciano del Corro[3], Johannes Hoffart[3], Gerhard Weikum[1]

July 18, 2018

[1]Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany
[2]Bosch Center for Artificial Intelligence, Renningen, Germany
[3]Ambiverse GmbH, Saarbrücken, Germany

Bush to Stress Domestic Issues in Speech.

Bush to Stress Domestic Issues in Speech.

George W. Bush

Bush to Stress Domestic Issues in Speech. *(Year 1989)*

George W. Bush

Bush to Stress Domestic Issues in Speech. *(Year 1989)*

George W. Bush

George H. W. Bush

Table of contents

Introduction

Problem Description

Given:

- Set of entity mentions $\mathcal{M}$ in a document.
- Entities: entries in a Knowledge Base (*KB*).

Task:

- Link each m, where $m \in \mathcal{M}$, to its correct entry in *KB*, if available.
- Predict as an *OOKBE*, otherwise.

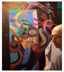
Named Entity Disambiguation

In 1959, **David Pearson** exhibited as part of the Young Contemporaries exhibition in **London**.

In 1981, with a small number of **BNR** colleagues, **David Pearson** left to found **Orcatech Inc.**

David Pearson raced for **Hoss Ellington** during the 1980 season.

Named Entity Disambiguation



(en.wikipedia.org/wiki/Dave_Pearson_(painter))

In 1959, **David Pearson** exhibited as part of the Young Contemporaries exhibition in **London**.

(en.wikipedia.org/wiki/David_Pearson_(computer_scientist))

In 1981, with a small number of **BNR** colleagues, **David Pearson** left to found **Orcatech Inc.**

David Pearson raced for **Hoss Ellington** during the 1980 season.

(en.wikipedia.org/wiki/David_Pearson_(racing_driver))

Context Evolution

- ## Popularity-based Models

Mihalcea and Csomai, 2007 [7]

Entity popularity and mention-entity prior probabilities.

Leverages anchor links structure.

David Pearson	`Dave Pearson (painter)`	0.1
David Pearson	`David Pearson (computer scientist)`	0.03

Context Evolution

Popularity-based Models

Mihalcea and Csomai, 2007 [7]

Entity popularity and mention-entity prior probabilities.

Leverages anchor links structure.

David Pearson	`Dave Pearson (painter)`	0.1
David Pearson	`David Pearson (computer scientist)`	0.03

Local Models

Bunescu and Pasca, 2006[2]; Cucerzan, 2007[3]; Milne and Witten, 2008[8]

Similarity with immediate context words.

Independent disambiguation.

David Pearson	`Dave Pearson (painter)`	1959, exhibited, young, exhibition, london
David Pearson	`David Pearson (computer scientist)`	1981, bnr, colleagues, found, orcatech

Context Evolution

- ## Global Models

  Kulkarni et al., 2007[6], Hoffart et al., 2011[4]

  Entities mentioned in a document are related.

  Collectively disambiguate entities.

David Pearson	`Dave Pearson (painter)`	*`London`*
David Pearson	`David Pearson (computer scientist)`	*`BNR, Orcatech_Inc.`*

Context Evolution

Global Models

Kulkarni et al., 2007[6], Hoffart et al., 2011[4]

Entities mentioned in a document are related.

Collectively disambiguate entities.

David Pearson	`Dave Pearson (painter)`	*London*
David Pearson	`David Pearson (computer scientist)`	*BNR, Orcatech_Inc.*

Representation Learning and Context Attention

Blanco et al., 2015[1], Hu et al.[5], 2015, Yamada et al, 2016[10]

Use of distributed vector representations.

Trained using the anchor links structure of KB.

Remove noisy words from the context.

David Pearson	`Dave Pearson (painter)`	$V_{London}, V_{exhibition}$
David Pearson	`David Pearson (computer scientist)`	$V_{BNR}, V_{Orcatech}$

Context Evolution

Temporal Context

Motivation for Temporal Modeling

Deductions

- Previous works fail to factor-in temporal semantics.
- Single value for entity popularity.
- Bias towards frequently occurring entities in KB and recent news.

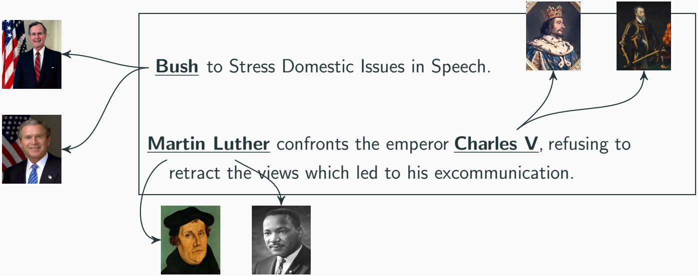


Figure 1: Entity Annotated Sample Texts[1]. (Image source: Wikipedia)

[1] The values in the brackets indicate the entity popularity.

Motivation for Temporal Modeling

Deductions

- Previous works fail to factor-in temporal semantics.
- Single value for entity popularity.
- Bias towards frequently occurring entities in KB and recent news.

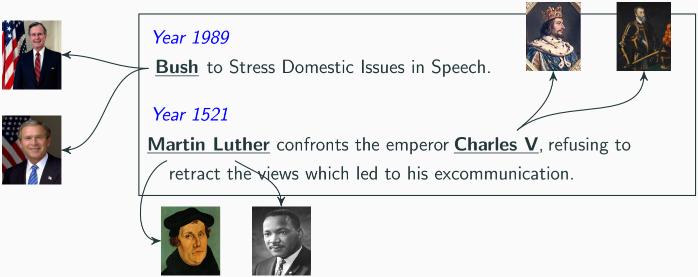


Figure 1: Entity Annotated Sample Texts[1]. (Image source: Wikipedia)

[1] The values in the brackets indicate the entity popularity.

Motivation for Temporal Modeling

Deductions

- Previous works fail to factor-in temporal semantics.
- Single value for entity popularity.
- Bias towards frequently occurring entities in KB and recent news.

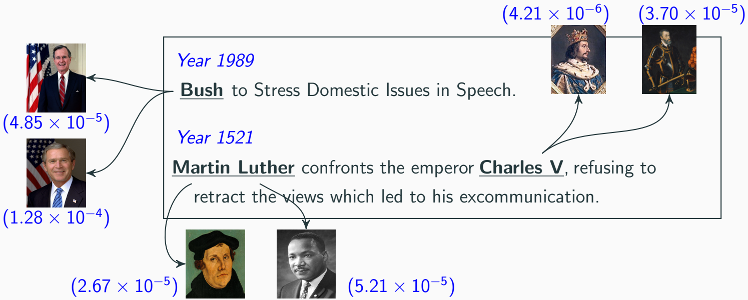


Figure 1: Entity Annotated Sample Texts[1]. (Image source: Wikipedia)

[1] The values in the brackets indicate the entity popularity.

Context Evolution

Temporal Context

Factor-in temporal semantics.

Distributed popularity.

Independent of anchor link structure.

Unbiased towards document creation time.

Temporal NED Model

Vector Space Modeling

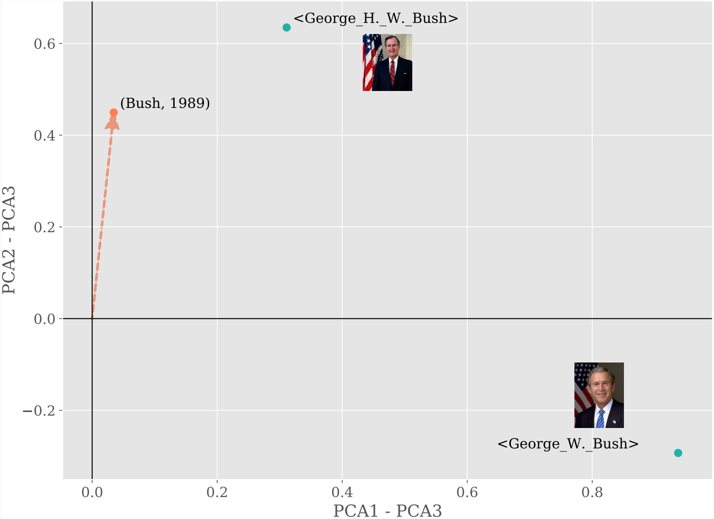


Figure 2: Temporal Vector Space Modeling[2].

[2]Representations: Entity as `<entity signature>` and mention as `(mention, year)`

Temporal Signatures of KB Entities

Martin Luther

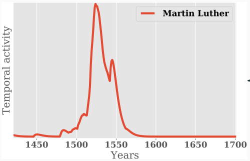
Martin Luther (**10 November 1483 18 February 1546**) was a German professor .. Luther proposed .. Ninety-five Theses of **1517** .. Leo X in **1520** and .. Diet of Worms **one year later** .. family moved to Mansfeld in **1484**, .. town councilor in **1492** .. Magdeburg in **1497** .. and Eisenach in **1498** .. In **1501**, at .. received his master's degree in **1505**.

HeidelTime

`1483-11-10, 1546-02-18, 1517, 1520, 1521, 1484, 1492, 1497, 1498, 1505`

(multi-set of temporal expressions)

Fix granularity → Exp. Smoothing



(signature)

Figure 3: Extraction of Temporal Signatures from Wikipedia Article Content.

Temporal Context for Entity Mentions

1. **Document Creation Time (DCT):** t_m^{dct}
   - Mention is represented as *One-Hot Vector*.
   - Applicable for news articles.
   - All values in the vector are 0, except a single 1 at the index position corresponding to *DCT*.
2. **In-context Temporal Information:** $t_m^{content}$
   - In-context expressions can be extracted using a temporal tagger.
   - Applicable for narrative documents.
   - There are 1s at index positions corresponding to the set of date values $\mathcal{T}(m)$ extracted by the temporal tagger.
3. **Combined Contexts:** t_m
   - The context similarity scores can also be aggregated.
   - $t_m = \lambda.t_m^{dct} + (1 - \lambda).t_m^{content}$

Disambiguation Example

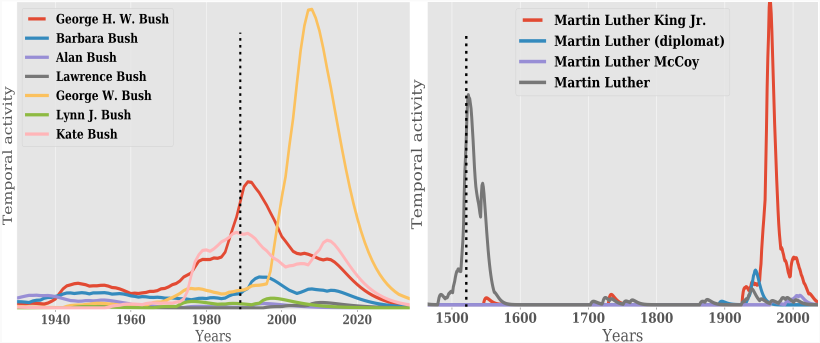


Figure 4: Temporal signatures of entity candidates for mentions *(Bush, 1989)* and *(Martin Luther, 1521)*.

Time-Aware Start-of-the-Arts

Making NEDs Time-aware

diaNED-1, extension of [Hoffart et al.: Robust Disambiguation of Named Entities in Text, EMNLP 2011]

- Document as a graph with mentions and entities as nodes. Mention-entity priors, mention entity similarity, and entity coherence used as edge weights.
- Disambiguation: A one-one mapping between each mention and entity node..

diaNED-2, extension of [Yamada et al.: Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation, SIGNLL 2016]

- Representation of context words and entities in a single vector space using skip gram model.
- Disambiguation: A learning-to-rank model using prior stats, string similarity, mention-entity, and coherence similarity as features.

Evaluation

Standard NED Datasets

CoNLL-AIDA	1996
TAC 2010	2004-2007
Microposts 2014	2011

Shortcomings

- Minimal improvements with Time-aware models.
- Not suitable to demonstrate/evaluate power of time-awareness.

A Diachronic Dataset: diaNED

HistoryNet

- *Historynet.com*: online resource of major historical events.
- Manually annotated 865 mentions in 350 randomly selected documents[3].

NewYorkTimes

- *NYT* headlines published between 1987 and 2007.
- Manually annotated 368 mentions in 300 randomly selected headlines.

[3] The named entities were identified using the 3 class Stanford NER tagger

Results: diaNED-1

	HistoryNet		NewYorkTimes	
Feature set	w/o time	w/ time	w/o time	w/ time
Prior	72.26	80.48*	38.14	54.24*
Context	63.63	66.10*	48.31	62.71*

Table 1: Micro-accuracy of diaNED-1 with and without time-awareness feature.

* significant over w/o time (Welch's t-test at level of 0.01).

Results: diaNED-2

	HistoryNet		NewYorkTimes	
Feature set	w/o time	w/ time	w/o time	w/ time
Base	89.44	90.23*	85.81	87.36*
String	89.40	90.00*	86.28	87.07*
Context	91.10	91.81*	87.07	88.34*
Coherence	91.16	91.98*	86.83	88.69*

Table 2: Micro-accuracy of diaNED-2 with and without time-awareness feature.

* significant over w/o time (Welch's t-test at level of 0.01).

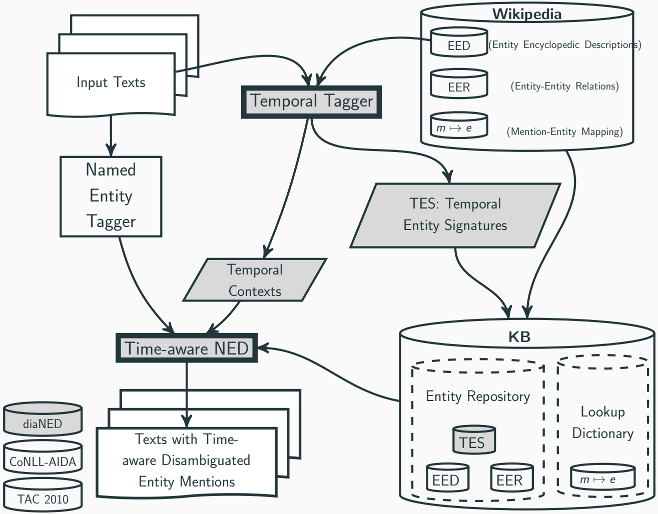
Results: diaNED

system	*HistoryNet*	*NewYorkTimes*
xLisa-NGRAM [Zhang and Rettinger, 2014]	87.07	66.30
xLisa-NER [Zhang and Rettinger, 2014]	83.32	60.25
WAT [Ferragina and Scaiella, 2012]	82.26	70.95
PBOH [Ganea et al., 2016]	90.26	71.75
FREME NER [Dojchinovski and Kliegr, 2013]	48.50	45.27
FRED [Consoli and Recupero, 2015]	23.18	15.44
FOX [Speck and Ngomo, 2014]	77.85	54.25
Dexter [Ceccarelli et al., 2013]	69.88	49.12
DBpedia Spotlight [Mendes et al., 2011]	56.92	61.91
AIDA [Hoffart et al, 2011]	82.68	70.14
AGDISTIS [usbeck et al, 2014]	70.77	50.14
Gupta et al., 2017	62.82	43.33
re-impl. of [Yamada et al., 2016]	90.87	72.55
diaNED-2	**91.68**	**76.09**

Table 3: Micro-f1 scores on the *HistoryNet* and *NewYorkTimes* datasets of diaNED-2 (trained on CoNLL-AIDA [4]) and other tools available on GERBIL [9].

Summary

Summary



Input Texts
Temporal Tagger
Wikipedia
EED
(Entity Encyclopedic Descriptions)
EER
(Entity-Entity Relations)
$m \mapsto e$
(Mention-Entity Mapping)
Named Entity Tagger
TES: Temporal Entity Signatures
Temporal Contexts
Time-aware NED
KB
Entity Repository
TES
EED
EER
Lookup Dictionary
$m \mapsto e$
diaNED
CoNLL-AIDA
TAC 2010
Texts with Time-aware Disambiguated Entity Mentions

Note

The annotated **diaNED Corpora** and **Entity Temporal Signatures** are available at: *https://www.mpi-inf.mpg.de/yago-naga/dianed/*

Future Work

- Study how temporal affinity can be used for identifying out-of-KB entities.
- Large scale experiments using data-sets generarted using semi-supervised methods.
- Adding multilingual support for the temporal signatures.

Thank you! Questions?

References i

R. Blanco, G. Ottaviano, and E. Meij.
Fast and space-efficient entity linking for queries.
In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 179–188. ACM, 2015.

R. C. Bunescu and M. Pașca.
Using encyclopedic knowledge for named entity disambiguation.
In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '06, pages 9–16, 2006.

References ii

S. Cucerzan.
Large-scale named entity disambiguation based on Wikipedia data.
In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '07, pages 708–716. Association for Computational Linguistics, June 2007.

J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum.
Robust disambiguation of named entities in text.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 782–792. Association for Computational Linguistics, 2011.

References iii

Z. Hu, P. Huang, Y. Deng, Y. Gao, and E. Xing.
Entity hierarchy embedding.
In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1292–1300, Beijing, China, July 2015. Association for Computational Linguistics.

S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti.
Collective annotation of wikipedia entities in web text.
In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 457–466. ACM, 2009.

References iv

R. Mihalcea and A. Csomai.
Wikify!: Linking documents to encyclopedic knowledge.
In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 233–242. ACM, 2007.

D. Milne and I. H. Witten.
Learning to link with wikipedia.
In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 509–518. ACM, 2008.

References v

R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann.
Gerbil: General entity annotator benchmarking framework.
In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1133–1143. International World Wide Web Conferences Steering Committee, 2015.

I. Yamada, H. Shindo, H. Takeda, and Y. Takefuji.
Joint learning of the embedding of words and entities for named entity disambiguation.
In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, CoNLL '15, pages 250–259. Association for Computational Linguistics, August 2016.