Web Science – **Investigating the Future** of Information and Communication

BlogNEER

Applying Named Entity Evolution Recognition on the Blogosphere

Helge Holzmann*, Nina Tahmasebi and Thomas Risse***
* L3S Research Center, {holzmann,risse}@L3S.de **Chalmers University of Technology, ninat@chalmers.se

Outline

- Introduction to Named Entity Evolution Recognition (NEER)
- Overview of BaseNEER on the New York Times
- Limitations of NEER on Blogs (vs. High Quality Newspaper)
- Approach (from BaseNEER to BlogNEER)
  - → Dataset Reduction
  - → Frequency Filtering
  - → Semantic Filtering
- Evaluation
- Conclusions

Language is Dynamic

- **Terms change over time**
- **Meanings change over time**
- **Different cultures lead to different language trends**
- **Local language trends spread globally on the Web**
- **Short living terms are preserved in digital archives**

- **Names of entities change over time**
  - → ***Joseph Ratztinger → Pope Benedict XVI***
  - → ***Czechoslovakia → Czech Republic, Slovakia***
  - → ***Sean Combs → Puff Daddy → P. Diddy***

Named Entity Evolution Recognition (NEER)

- **Detection of name changes and alternative names**
  - **→ Temporal co-references**
    - Direct, e.g., *Barack Obama ↔ President Obama* *(lexical overlap)*
    - *Indirect, e.g., Project Natal ↔ Kinect* *(no lexical overlap)*
- **Support for information retrieval**
  - **→ Especially on datasets covering long time ranges (digital archives)**
    - Query expansion:

BaseNEER *

NEER: An Unsupervised Method for Named Entity Evolution Recognition*

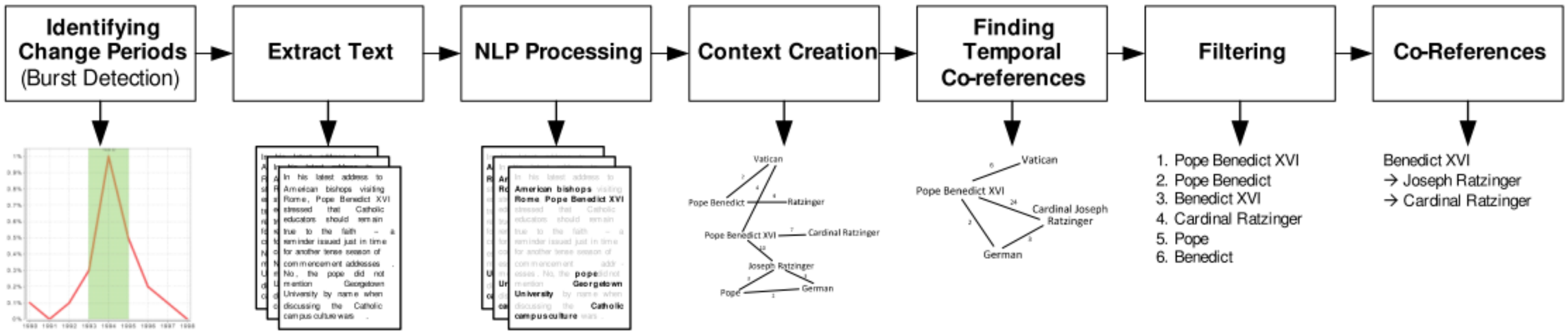
Nina TAHMASEBI *Gerhard GOSSEN* *Nattiya KANHABUA*
Helge HOLZMANN *Thomas RISSE*
L3S Research Center, Appelstr. 9a, 30167 Hannover, Germany
{tahmasebi, gossen, kanhabua, holzmann, risse}@L3S.de

ABSTRACT
High impact events, political changes and new technologies are reflected in our language and lead to constant evolution of terms, expressions and names. Not knowing about names used in the past for referring to a named entity can severely decrease the performance of many computational linguistic algorithms. We propose NEER, an unsupervised method for named entity evolution recognition independent of external knowledge sources. We find time periods with high likelihood of evolution. By analyzing only these time periods using a sliding window co-occurrence method we capture evolving terms in the same context. We thus avoid comparing terms from widely different periods in time and overcome a severe limitation of existing methods for named entity evolution, as shown by the high recall of 90% on the New York Times corpus. We compare several relatedness measures for filtering to improve precision. Furthermore, using machine learning with minimal supervision improves precision to 94%.

** 24th International Conference on Computational Linguistics (Coling 2012) Mumbai, India, December 2012 http://www.l3s.de/neer-dataset*

BaseNEER

"Chad Johnson has legally changed his name to Chad Javon Ocho Cinco"
[sports.espn.go.com]



NEER on Blog Data vs. High Quality Newspaper

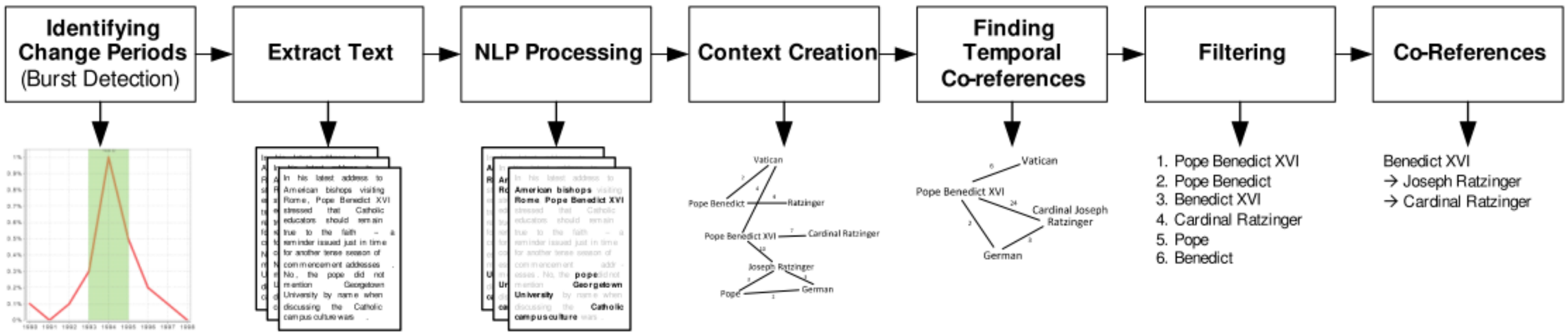
- Multiple sources vs. one source
- More dynamic language vs. editorial controlled
- Rather colloquial vs. written/formal
- Linking complementary terms/entities vs. focused reports

→ More co-occurring terms

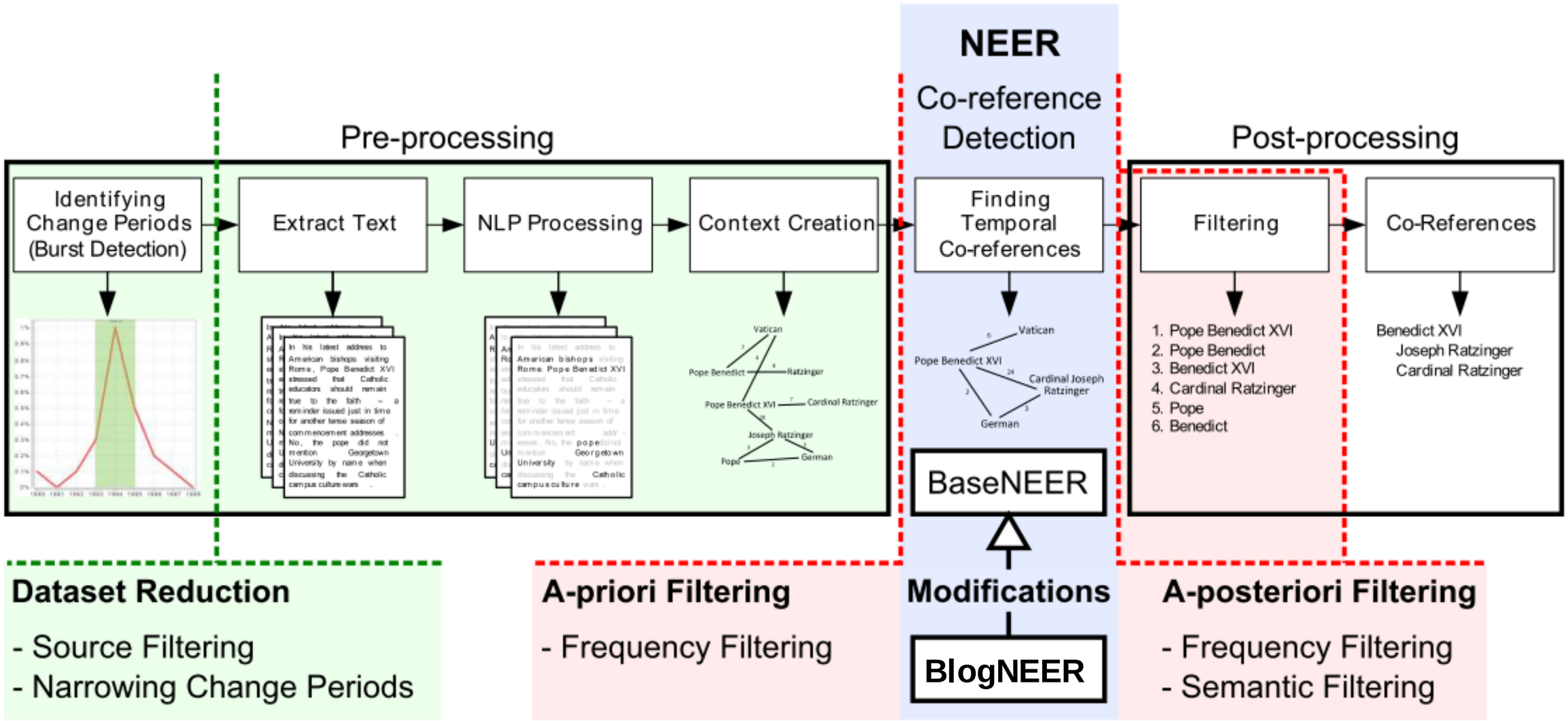
→ Larger contexts

→ More noise

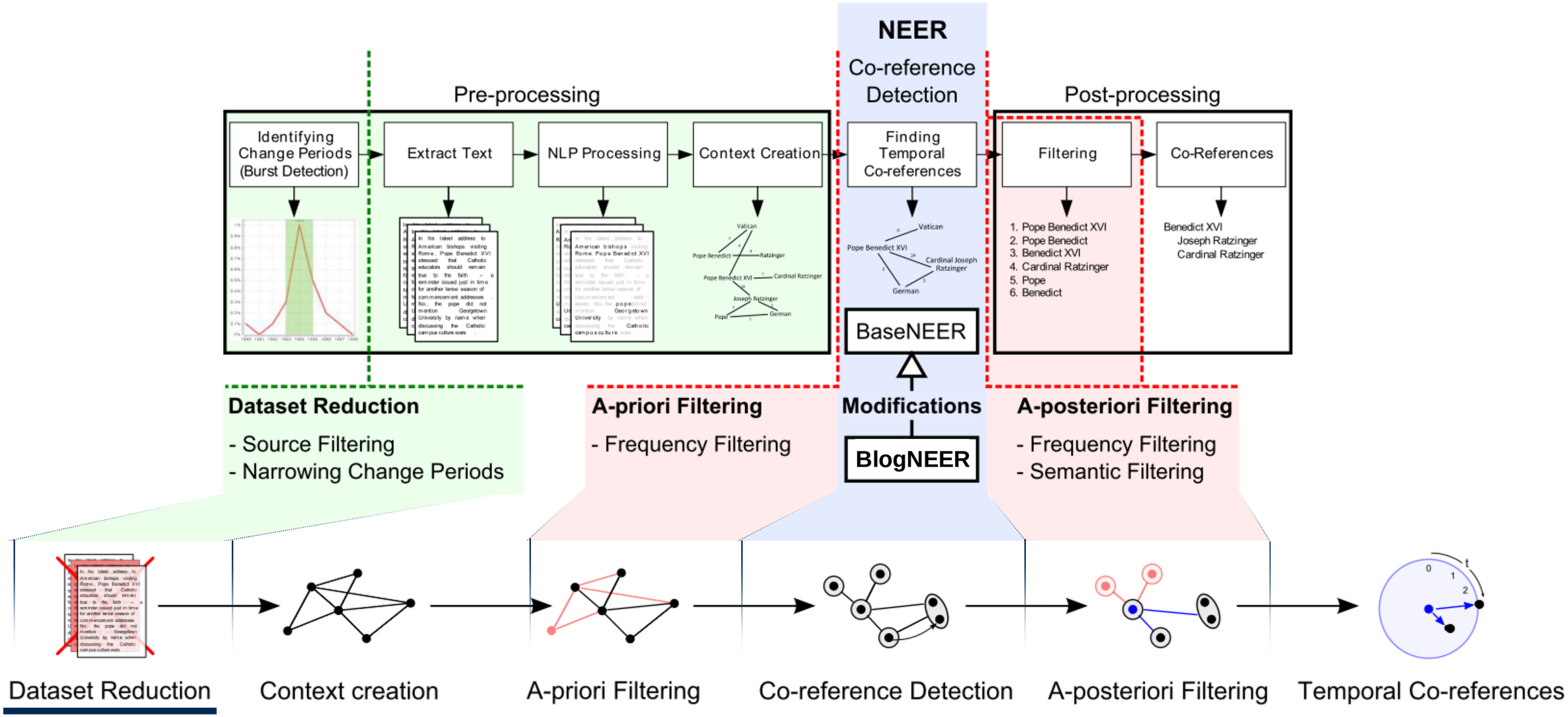
From BaseNEER to BlogNEER



From BaseNEER to BlogNEER



From BaseNEER to BlogNEER



BlogNEER workflow

Dataset Reduction

- **Source filtering**

  → Consider *President Obama* and a president of some sports club

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Okt	Nov	Dec
Source 1			▯ ▯			▯		▯ ▯	▯			
Source 2		▯ ▯				▯		▯		▯	▯	▯
Source 3							▯	▯	▯	▯	▯	▯
Source 4			▯	▯	▯ ▯	▯ ▯ ▯	▯ ▯ ▯	▯ ▯ ▯	▯ ▯	▯ ▯		
Source 5			▯					▯ ▯	▯	▯ ▯ ▯		

▯ Document containing *President* **or** *Obama*

Dataset Reduction

- **Source filtering**

  → Consider *President Obama* and a president of some sports club

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Okt	Nov	Dec
Source 1												
Source 2												
Source 3												
Source 4												
Source 5												

Document containing *President* **or** *Obama* Document containing *President Obama*

Dataset Reduction

- **Narrowing change periods**

  → Consider the presidential election and other events involving Obama

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Okt	Nov	Dec
Source 1												
Source 2												
Source 3												
Source 4												
Source 5												

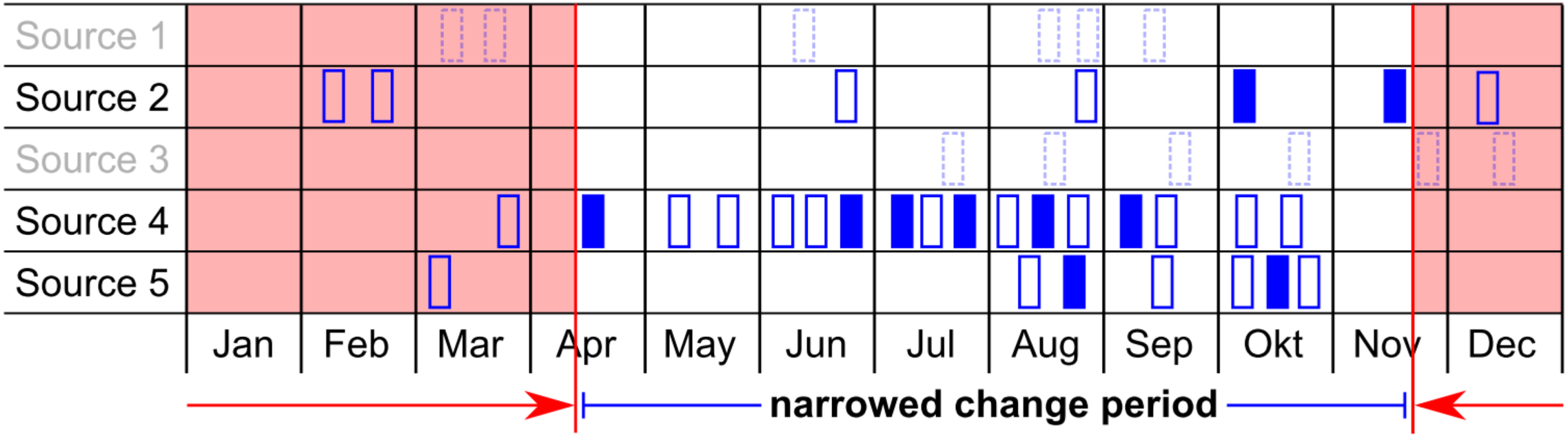
Document containing *President* **or** *Obama*

Document containing *President Obama*

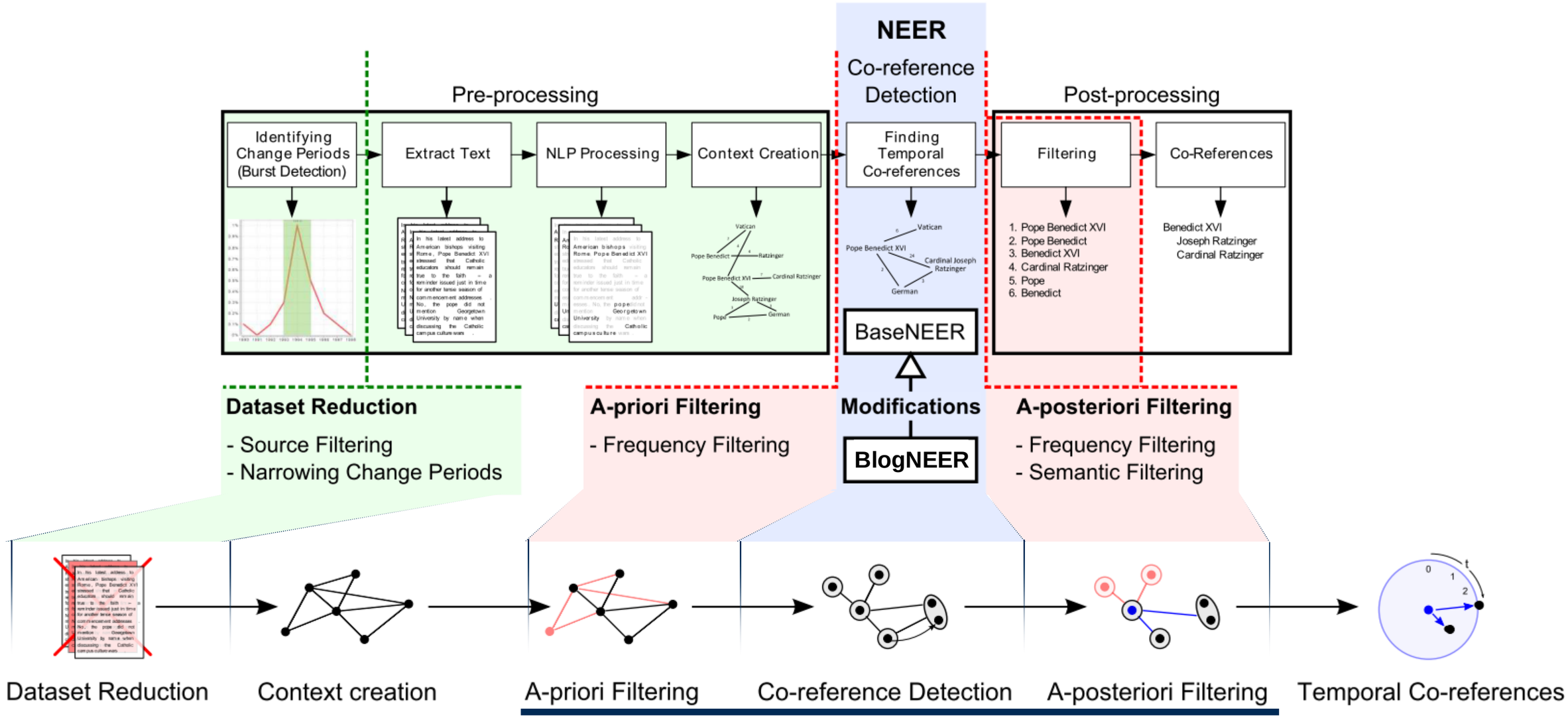
Dataset Reduction

- **Narrowing change periods**

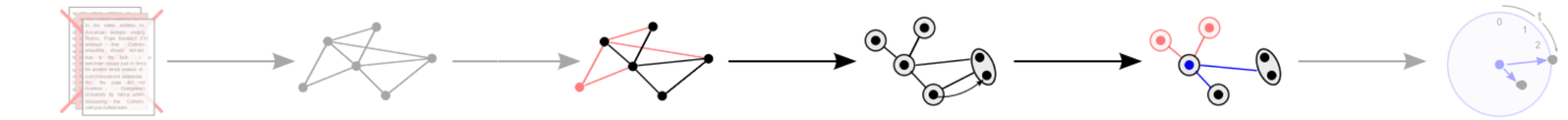
  → Consider the presidential election and other events involving Obama



Frequency Filtering and Co-reference Detection

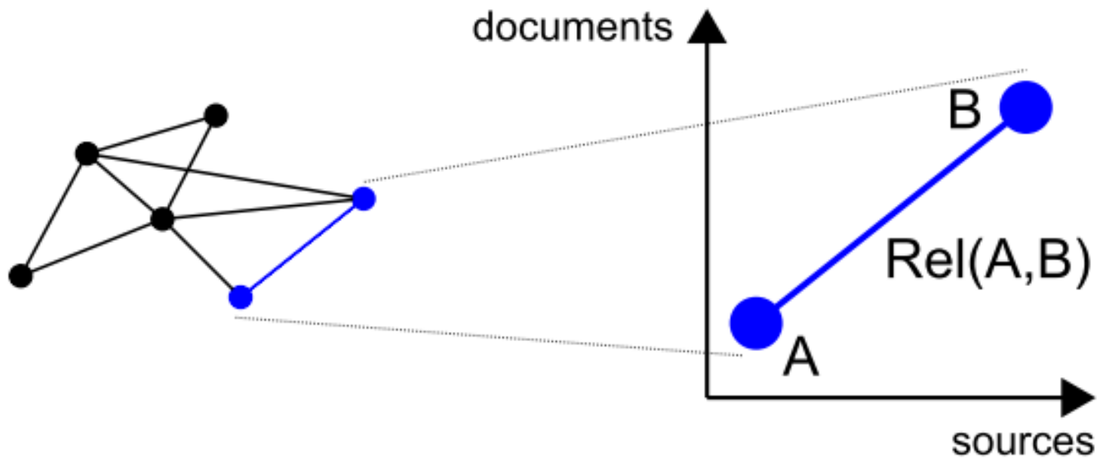


BlogNEER workflow



Frequency Filtering and Co-reference Detection

- **Based on number of documents / sources of terms / relations**
  - → **E.g.,** *"USA Today just leaked the name of Microsoft's Project Natal motion control setup: Kinect"* [engadget.com]



- **Co-reference detection**
  - → **Term merging / graph consolidation by means of sub-terms**
    - E.g., *Tony Blair* ↔ *Prime Minister Tony Blair* ↔ *Prime Minister*
  - → **A-posteriori frequency filtering based on accumulated frequencies**

Semantic Filtering

- **Incorporating DBpedia**

  → http://dbpedia.org/resource/Pope_Benedict_XVI

About: Pope Benedict XVI

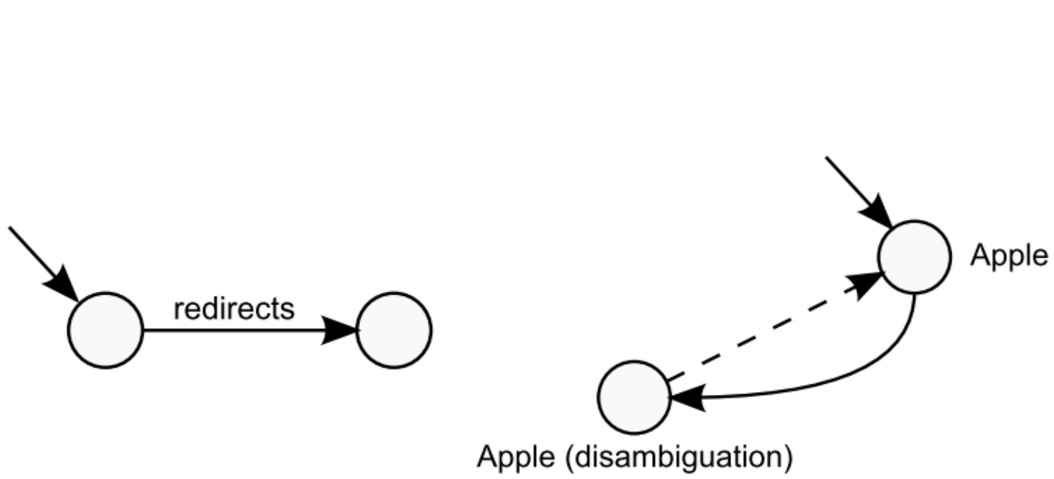
An Entity of Type : ChristianBishop, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

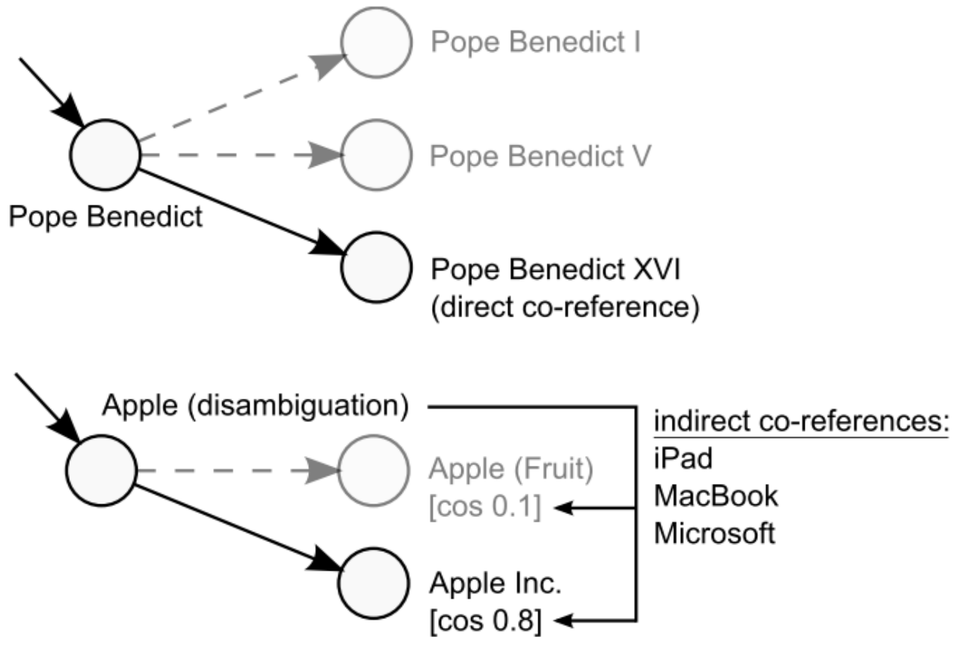
Benedict XVI (Latin: Benedictus PP. XVI; Italian: Benedetto XVI; German: Benedikt XVI. ; born Joseph Aloisius Ratzinger on 16 April 1927) is the 265th Pope, a position in which he serves dual roles as Sovereign of the Vatican City State and leader of the Catholic Church. As pope he is regarded as the successor of Saint Peter.

Property	Value
dbpedia-owl:birthDate	▪ 1927-04-16 (xsd:date)
dcterms:subject	▪ ▪ ▪ ▪ category:German_Roman_Catholic_theologians ▪ category:German_popes ▪ category:Grand_Crosses_of_the_Order_of_Merit_of_the_Federal_Republic_of_Germany ▪ ▪ ▪
rdf:type	▪ owl:Thing ▪ dbpedia-owl:Agent ▪ dbpedia-owl:Person ▪ ▪ ▪
foaf:name	▪ Pope Benedict XVI
is dbpedia-owl:wikiPageDisambiguates of	▪ dbpedia:Pope_(disambiguation) ▪ dbpedia:Pope_Benedict ▪ dbpedia:Benedict
is dbpedia-owl:wikiPageRedirects of	▪ dbpedia:Benedict_XVI ▪ ▪ ▪

Semantic Filtering

- **Incorporating DBpedia**
  - **→ Disambiguation and aggregation of properties**
    - Following redirections
    - Redirecting to a disambiguation resource
    - Disambiguation by means of direct/indirect co-references

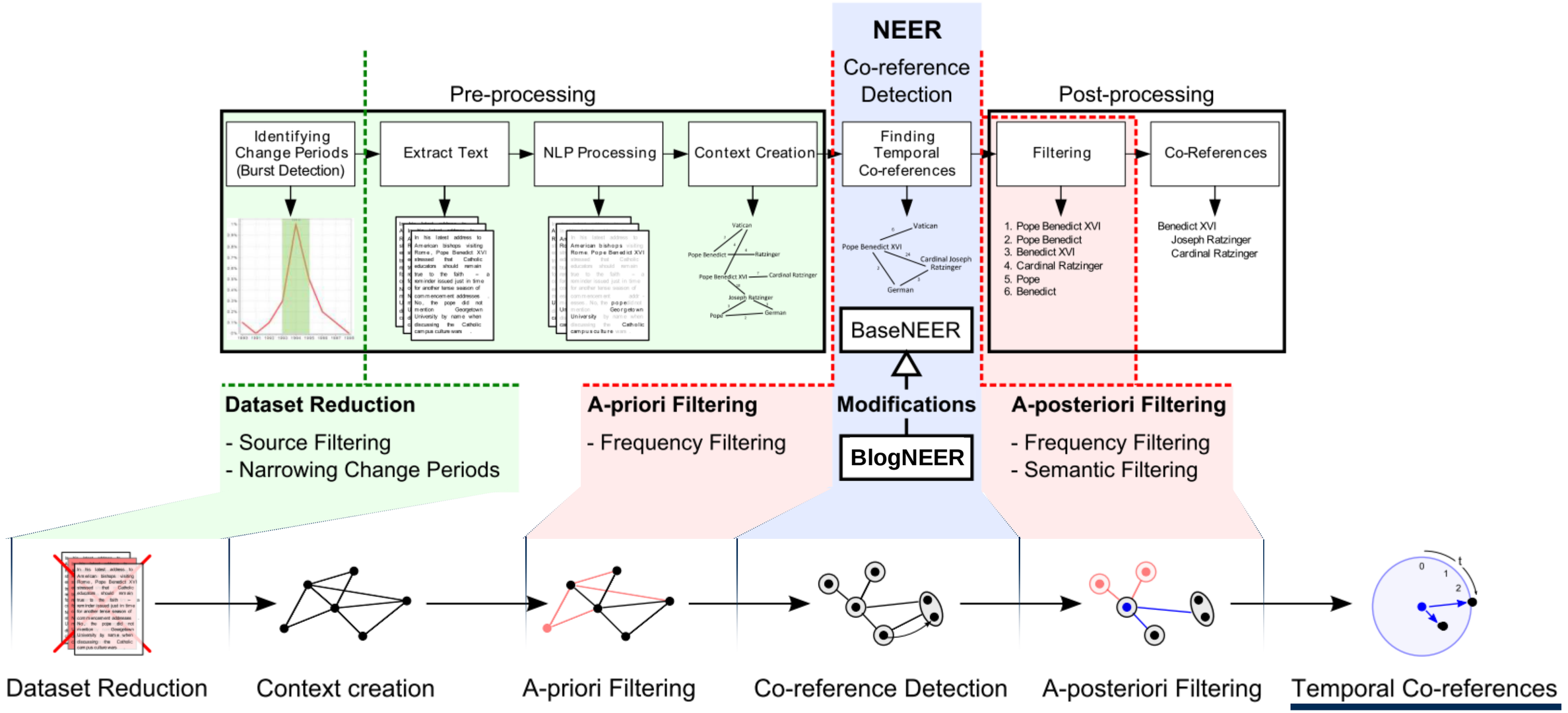




Semantic Filtering

- **Comparing detected co-reference candidates**
  - → **Incorporating semantic properties**
    - Intersections of type, subject, year sets
    - Type hierarchy comparison
  - → **Consider: *Pope Benedict XVI* vs. *Vatican***
    - Types: Person vs. Place
    - Subjects: German popes vs. Holy cities
    - Years: 1927 (birth date) vs. 1992 (founding date)

Evaluation



BlogNEER workflow

Evaluation

- **NEERfx experimental framework**

- **Two blog datasets**
  - → **Blogs08** *
    - English texts from first 10% of TREC-2008 blog dataset
  - → **Technorati** **
    - Top 100 blogs of nine categories parsed from 2005 to 2013
- **BaseNEER test set**
- **Performance measures: precision > recall**

** Iadh Ounis, Craig Macdonald and Ian Soboroff. Overview of the trec-2008 blog track. In In Proceedings of TREC-2008, 2009.*
*** http://www.technorati.com*

Evaluation

- **BaseNEER vs. BlogNEER**

	Precision	Recall
BaseNEER	8%	90%
BaseNEER + frequency filtering	33%	86%
BlogNEER without a-posteriori filtering	6%	64%
BlogNEER after a-posteriori frequency filtering	48%	43%
BlogNEER after semantic filtering	67%	36%

Results on Blogs08

	Precision	Recall
BaseNEER	8%	90%
BaseNEER + frequency filtering	33%	86%
BlogNEER without a-posteriori filtering	6%	87%
BlogNEER after a-posteriori frequency filtering	61%	77%
BlogNEER after semantic filtering	70%	67%

Results on Technorati

Evaluation

- **Example**

Step	Result set	Precision	Recall
Unfiltered	*Apple, Engadget, GameStop, Project Natal, Kotaku, Nintendo, Redmond, USA Today, Microsoft Kinect, Microsoft*	20%	100%
Semantic Filtering	*Project Natal, Microsoft Kinect*	100%	100%

Query Kinect on Technorati

→ Expected *Project Natal* and *Microsoft Kinect*

Evaluation

- **Example**

Step	Result set	Precision	Recall
Unfiltered	*Sean, Sean Penn, Penn, Combs, Diddy, York, Puff, Puff Daddy, Daddy, MTV, Video, Video Music Awards, Music Awards, Music, Award, Boy, Rock, Chris Rock, Chris, Bad, Rapper, …*	12%	100%
Frequency Filtering	*Sean, Sean Penn, Combs, Diddy, Puff Daddy, Video Music Awards*	67%	100%
Semantic Filtering	*Puff Daddy*	100%	50%

Query Sean Combs on Blogs08

→ Expected *Puff Daddy* and *Diddy*

→ *Diddy* disambiguated to *Diddy – Dirty Money* ≠ *Sean Combs*

Conclusions

- **BlogNEER more resistant against noise (compared to BaseNEER)**
  - **→ Comparable / better results**
- **Dataset reduction very effective**
  - **→ Room for improvement, e.g., clustering of sources / documents**
- **First approach incorporating semantic filtering**
  - **→ Very promising results, also co-references unknown by DBpedia**
    - e.g., Czechoslovakia, Czech Republic, Slovakia
  - **→ Works only with available entities (*Diddy* example)**
- **Future work to focus on incorporating …**
  - **→ … semantic meta data (e.g., given name → gender)**
  - **→ … explicit temporal / co-reference information instead of co-occurrences**

Questions / Discussion

Thank you for your attention!