Entity and Aspect Extraction for Organizing News Comments

Radityo Eko Prasojo, Mouna Kacimi & Werner Nutt

EMCL Workshop, 11–12 February 2016

Comments in News Website

Typically, comments are listed based on date-time and reply relation.

Problem: difficulty to catch the flow of the discussions and to understand their main points of agreement and disagreement.

Example: why is independence good/bad for the Scots? Will their economy be affected?

There is a need for organizing comments

to help users to:

(1) have a better understanding of the viewpoints related to each topic
(2) facilitate the participation in discussions and thus increase the chance of acquiring new viewpoints

by clustering comments containing **similar discussions**:

- they talk about **the same entities**
- they argue about **the same aspects** of those entities.

Contributions

- Improvements on state-of-the-art unsupervised **entity extraction** tools (Zemanta, NERD, AIDA Yago)
  - Addressed issues: noises and low coverage (due to coreferences)
- Introduced **aspect extraction** in news domain
  - Previously: aspects only on product review domain (Zhang & Liu, 2014)

Entity extraction Baseline: Unsupervised Tools

"Don't be afraid of Rasmussen or NATO, this is none of their business. If he ends up sending USS Ronald Reagan aircraft carrier to the coast of Scotland, then he should have done the same to Crimea."

Zemanta

"Don't be afraid of Rasmussen or NATO, this is none of their business. If he ends up sending USS Ronald Reagan aircraft carrier to the coast of Scotland, then he should have done the same to Crimea."

<http://dbpedia.org/resource/NATO>

<http://dbpedia.org/resource/Crimea>

<http://dbpedia.org/resource/Aircraft_Carrier>

<http://dbpedia.org/resource/Scotland>

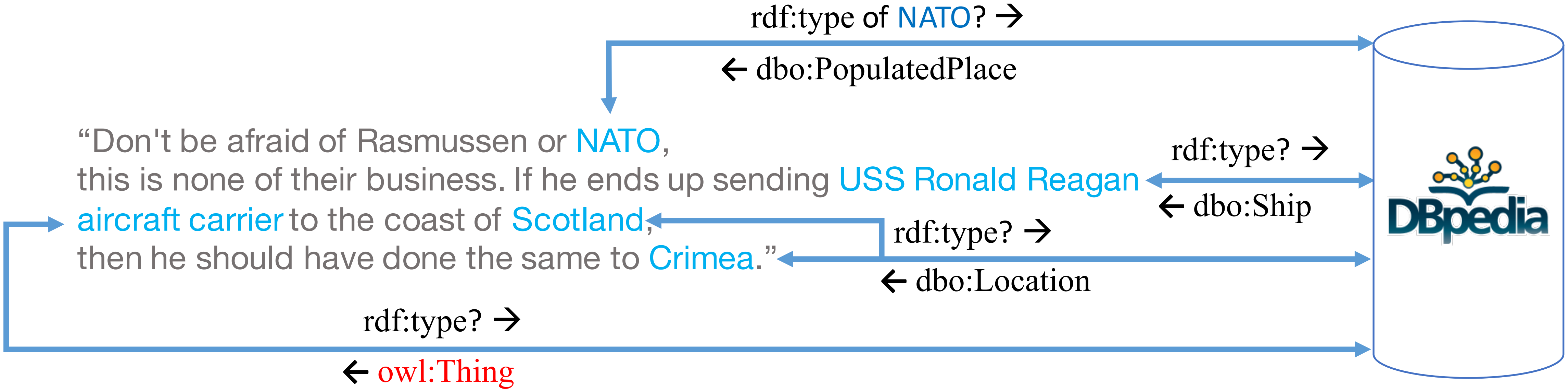
<http://dbpedia.org/resource/USS_Ronald_Reagan>

- Improved by applying:
  - Entity filtering
  - Name normalization
  - Entity search on KB
  - Coreference resolution

Tools: Dbpedia
Stanford CoreNLP

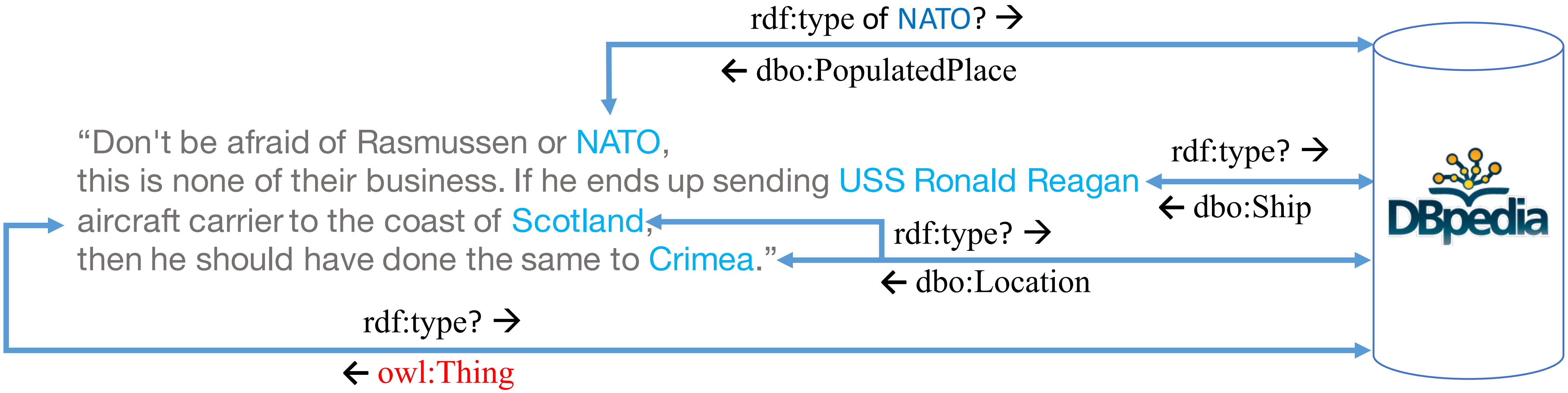
Entity Filtering

An entity is an **instance of some well-defined class**



Entity Filtering

An entity is an **instance of some well-defined class**



We remove entities that don't have rdf:type other than owl:Thing and owl:Class

Risk: lower recall

Name Normalization

An entity may appear using non-proper names (alias)

Scotland Rejects Independence in Referendum

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut suscipit lacus ac bibendum mollis. Pellentesque tincidunt vulputate ligula a efficitur. Aliquam erat

Anders Fogh Rasmussen

foaf:givenName,foaf:surname?→

←{'Anders', 'Fogh', 'Rasmussen'}

....

United States

dbpedia-owl:wikiPageRedirects of? →

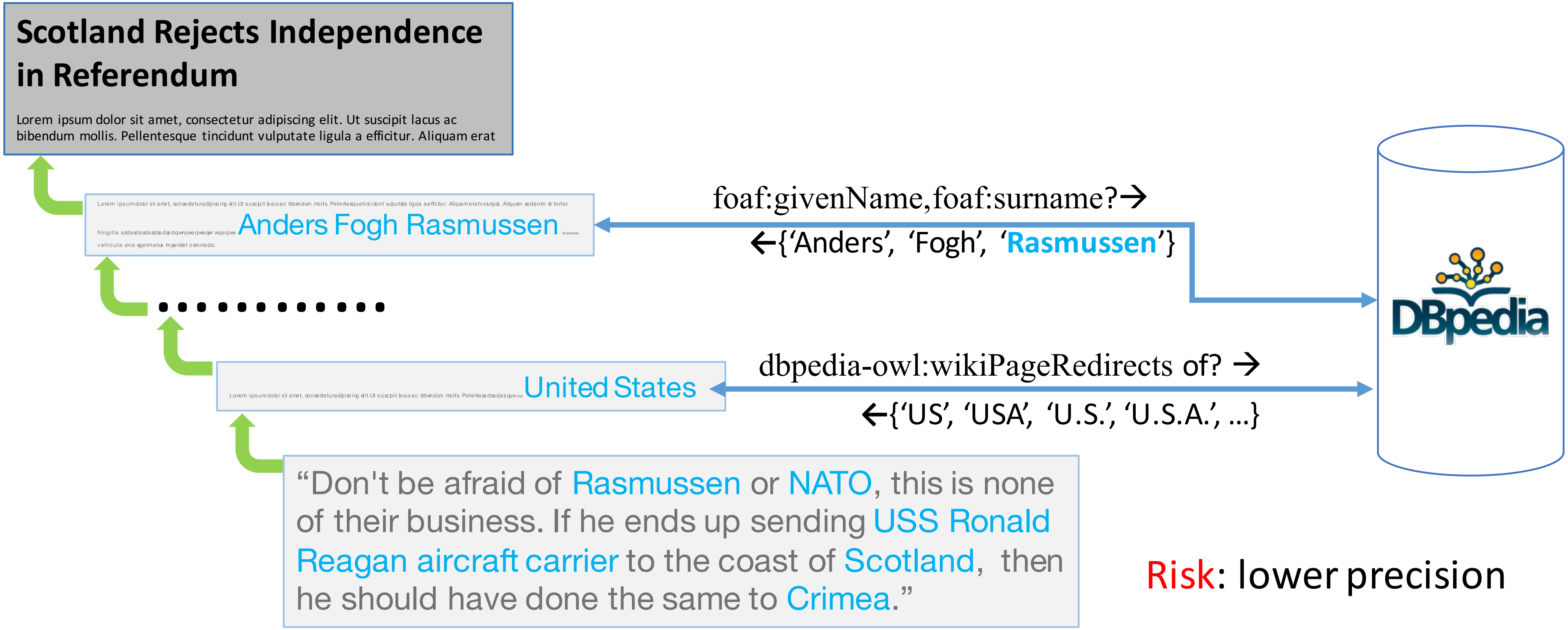
←{'US', 'USA', 'U.S.', 'U.S.A.', ...}

DBpedia

"Don't be afraid of Rasmussen or NATO, this is none of their business. If he ends up sending USS Ronald Reagan aircraft carrier to the coast of Scotland, then he should have done the same to Crimea."

Name Normalization

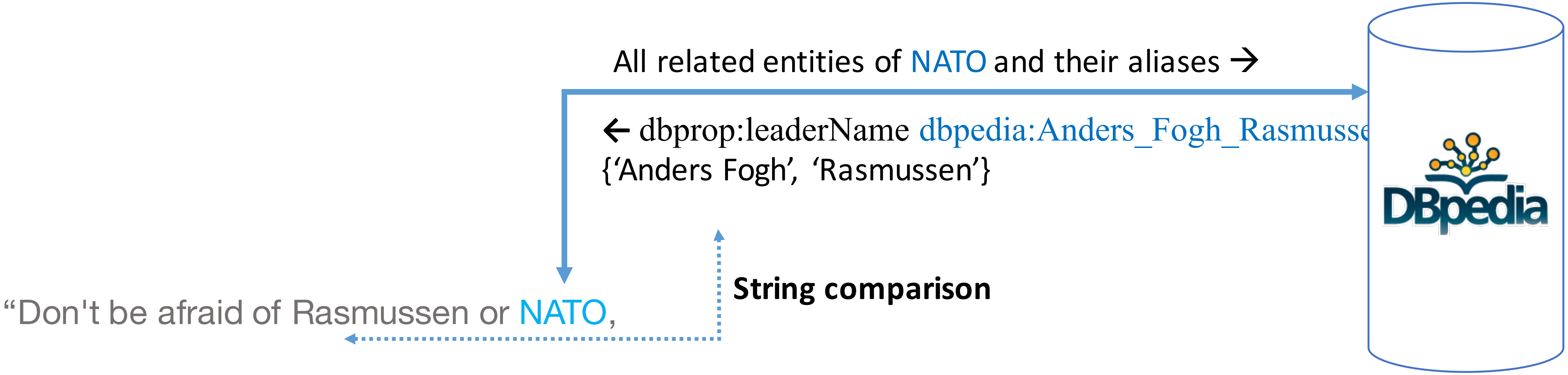
An entity may appear using non-proper names (alias)



Risk: lower precision

Context-Related Entity Search

Sometimes, mapping for an alias cannot be found



Context-Related Entity Search

Sometimes, mapping for an alias cannot be found

All related entities of NATO and their aliases →

← dbprop:leaderName dbpedia:Anders_Fogh_Rasmusse

{'Anders Fogh', '**Rasmussen**'}

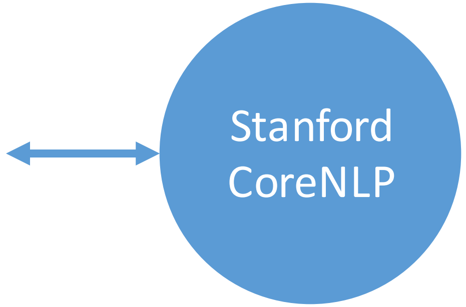
String comparison

"Don't be afraid of Rasmussen or NATO,

Risk: lower precision

Coreference Resolution

"Don't be afraid of Rasmussen or NATO, this is none of their business. If he ends up sending USS Ronald Reagan aircraft carrier to the coast of Scotland, then he should have done the same to Crimea."



Coreference Resolution

Coreference resolution

"Don't be afraid of Rasmussen or NATO, this is none of their business. If he ends up sending USS Ronald Reagan aircraft carrier to the coast of Scotland, then he should have done the same to Crimea."

Stanford CoreNLP

Risk: lower precision

Entity Extraction – Experiment Setup

- **10 news articles** that use DISQUS
- **100 comments** having the highest word counts
- **5 students** as **entity** and **aspect** annotators
- Annotated data as ground truth

Entity Extraction – Experiment Results

	Zemanta (baseline)		+Entity Filtering		+Name Normalization		+Context Search		+Coreference Resolution	
	Prec.	*Recall*	*Prec.*	*Recall*	*Prec.*	*Recall*	*Prec.*	*Recall*	*Prec.*	*Recall*
Politics	70.01	59.17	89.33	58.83	89.23	67.43	89.22	71.99	89.07	81.31
Techs	74.10	52.33	94.52	52.31	94.51	53.09	94.51	54.98	89.43	61.62
Sport	75.04	36.61	96.11	36.61	96.09	40.68	95.89	70.61	92.08	79.81
Average.	72.74	50.35	92.92	50.21	92.88	55.06	92.81	66.75	90.09	74.94

	AIDA (baseline)		+Entity Filtering		+Name Normalization		+Context Search		+Coreference Resolution	
	Prec.	*Recall*	*Prec.*	*Recall*	*Prec.*	*Recall*	*Prec.*	*Recall*	*Prec.*	*Recall*
Politics	77.82	66.35	78.08	66.34	80.21	67.88	80.17	68.51	80.01	77.56
Techs	83.51	69.33	83.75	69.33	88.37	75.75	87.33	76.07	83.30	85.98
Sport	91.55	31.97	91.89	31.95	91.89	32.44	90.86	45.64	86.67	51.74
Average	81.67	56.93	81.93	56.92	84.73	59.61	84.19	63.92	82.08	72.34

Aspects

- of product entities
  - *The aspects of an entity **e** are the components and attributes of **e**.* (Zhang & Liu, 2014)
- of entities on news
  - "More Scots would definitely have voted no than yes" (voting - **action**)
  - "Orlando Bloom is a good actor" (acting - **skill**)
  - "…it's the right the right of Scottish People" (right - **possession**)
  - Other: **components, attributes, and moods**

An **aspect** is all what is arguable about an **entity**

Types of Aspects

aspect: right

- "…it's the right the right of Scottish People." **(explicit)**

aspect: employer

- "Tesco is large." **(implicit)**

aspect: beauty

- "Scotland is a very beautiful country." **(semi-implicit)**

aspect: voting

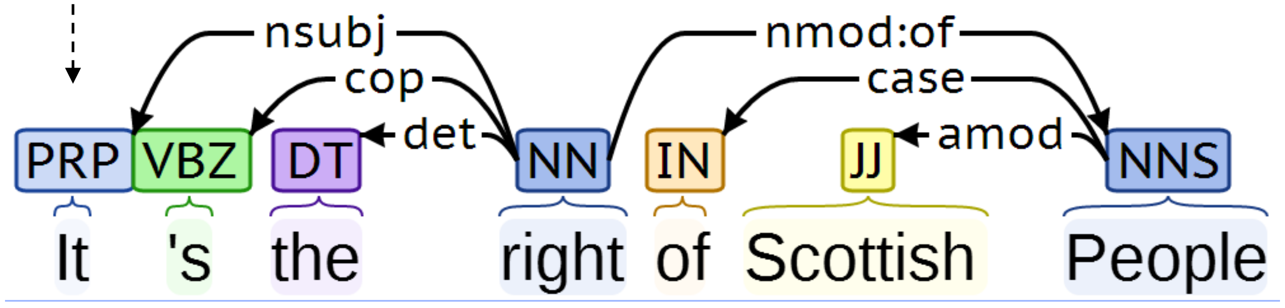
- "The Scots can vote however they want." **(semi-implicit)**

Extraction of Explicit Aspects: Exploiting Dependency

Prepositional Dependency

"...it's the right the right of Scottish People."

Stanford CoreNLP Annotation



Extract all **noun phrases** that have an **nmod:of**, **nmod:in**, **nmod:on**, or **nmod:at** relation towards the entity in the sentence.

Extraction of Explicit Aspects, Combined with Entity Extraction

- Specifically, the coreference resolution part

"Don't be afraid of Rasmussen or NATO, this is none of their business."

possessive dependency

Extraction of Implicit Aspects: Adjective-to-aspect Mapping

Idea from (Zhang & Liu, 2014)

"Tesco is large"

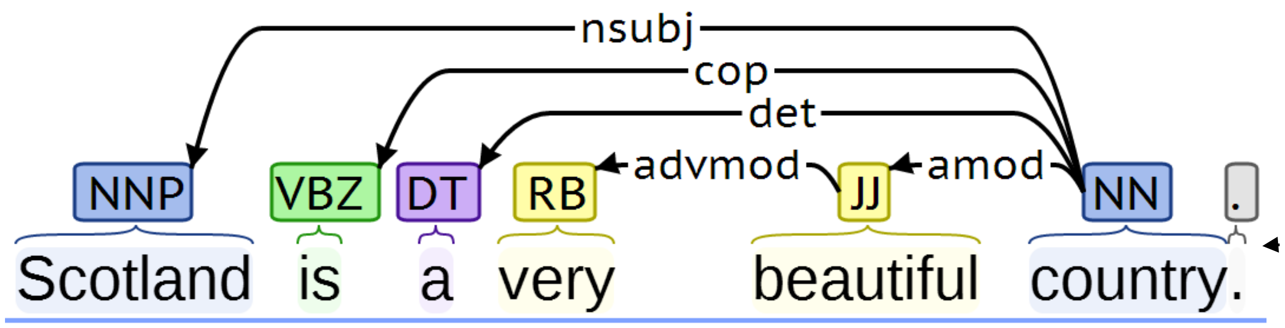
In other comments, we found as aspects of Tesco, qualified as "large":

- employer (2x)
- back office (1x)
- call center operation (1x)

We conclude: most probably, the employer aspect of Tesco was meant. Result is further improved by (1) taking into account frequent context words and (2) lexical relations of adjective.

Extraction of Semi-Implicit Aspects (1)

- Semi-implicit aspects: implicit aspects that don't have mapping.
- "Scotland is a very beautiful country."



Stanford CoreNLP Annotation

Lexical database search

beautiful $\xrightarrow{\text{WordNet:attribute}}$ beauty

We consider following lexical relations:
attribute > pertainym > participle of verb > derivationally related form (drf) > see also

We search for a **noun phrase** that is connected to the entity and the word indicating the aspect using a lexical relation.

Extraction of Semi-Implicit Aspects (2)

- Generally can be used to identify aspects from verb, adjective, or noun.
- Other examples:

beautiful $\xrightarrow{\text{WordNet:attribute}}$ beauty

instrumental $\xrightarrow{\text{WordNet:pertainym}}$ instrument

vote $\xrightarrow{\text{WordNet:derivationally related form}}$ voting

inexcusable $\xrightarrow{\text{WordNet:antonym}}$ excusable $\xrightarrow{\text{WordNet:similar to}}$ justifiable

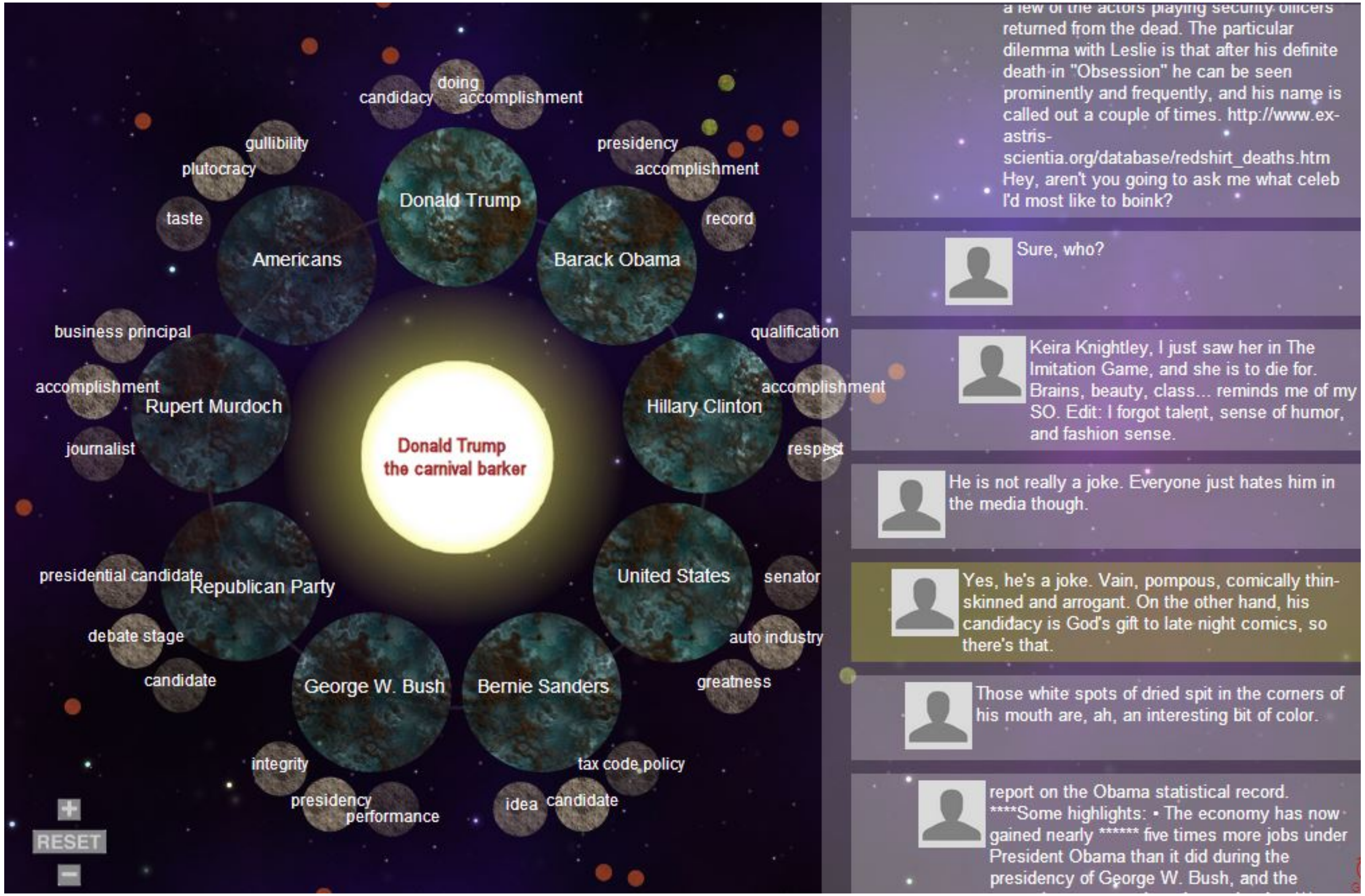
justification $\xleftarrow{\text{WordNet:drf}}$ justify $\xleftarrow{\text{WordNet:drf}}$ justifiable

- If there are multiple possible aspects for a single word, we use **WordNet::Similarity** to decide for the best one.

Aspect Extraction – Experiment and Results

	Precision	Recall	F_1
Explicit	**90.88**	23.50	37.34
Explicit + Implicit	76.87	26.12	38.99
Explicit + Semi-Implicit	71.31	64.62	67.80
Explicit + Implicit+ Semi-Implicit	73.12	**73.82**	**73.47**

Visualization



RE Prasojo, M Kacimi, F Darari. IEEE InfoVis. Chicago, 25-30 October 2015.

demo is now available at **orcaestra.inf.unibz.it**

Conclusion

- General contribution: a framework for organizing news comment using **entity extraction** and **aspect extraction**.

- Our **entity extraction**: unsupervised tools + **entity filtering**, **name normalization**, **entity search**, and **coreference resolution**.

- We extract **explicit**, **implicit**, and **semi-implicit** aspects using **grammar analysis** and **lexical database search**.

- Experiment shows improvement on both **entity** and **aspect** extraction compared to baseline technique.

Future Works

- Addressing limitations:
  - **Concept extractions**
  - Idioms and other metaphorical expressions
  - Difficult coreference (e.g. demonstrative pronouns)
  - Experiments on more, various data

- Complete the missing pieces:
  - **Sentiment analysis** for news comments

Acknowledgements

- European Master's Programme in Computational Logic
- To-Know Project
- SIGIR Student Travel Grant

Thank you!

Extra slides

An **entity** can be...

a person, a location, an organization, or **any well-defined concept** such as nationalities, languages, or wars.

Entity Extraction Tasks

1. **Recognition** – through proper names/rigid designators (Coates-Stephens, 1992) (Thielen, 1995) (Nadeau & Sekine, 2006)

   "**Scotland** can vote however it wants, it's the **Scottish peoples** right."

2. **Disambiguation** – by mapping to a representation in a KB

   <http://dbpedia.org/resource/USS_Ronald_Reagan_(CVN-76)>

   "If he ends up sending **USS Ronald Reagan** aircraft carrier to the coast of **Scotland**, then he should have done the same to **Crimea**."

   <http://dbpedia.org/resource/Scotland>

   <http://dbpedia.org/resource/Crimea>

Supervised vs Unsupervised Approaches to Entity Extraction

Aspect of **EE**	Supervised	Unsupervised
Prominent tools	StanfordNLP	Zemanta, AlchemyAPI, NERD, Aida YAGO
Recognition ability	depends on the training set	domain-independent
Disambiguation ability	limited	provided by KBs
Running time	fast	slow

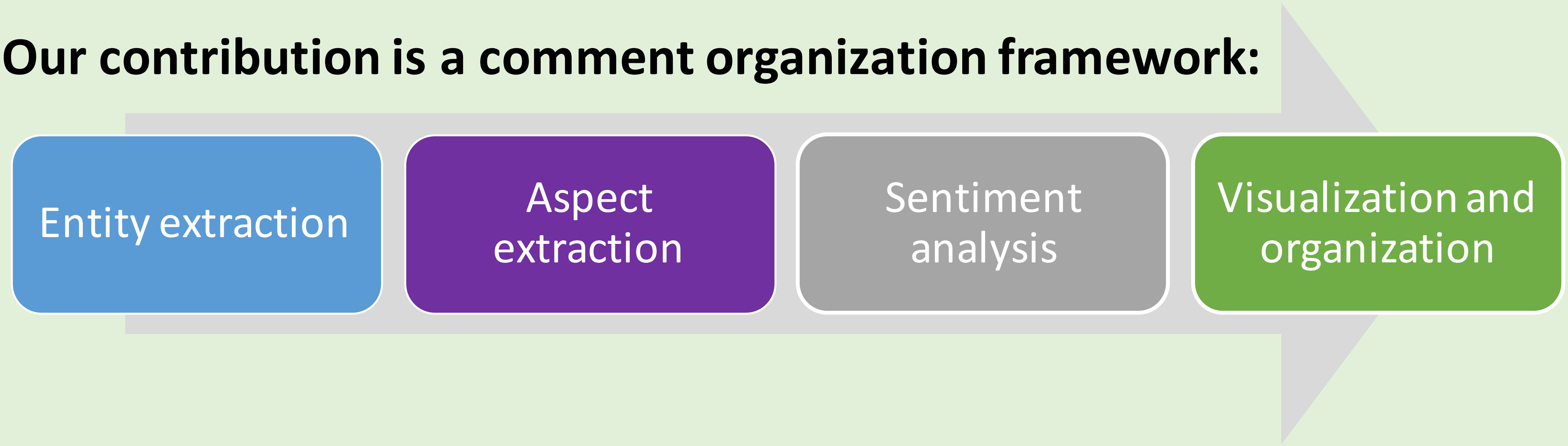
There is a need for organizing comments

to help users to:

(1) have a better understanding of the viewpoints related to each topic

(2) facilitate the participation in discussions and thus increase the chance of acquiring new viewpoints

Our contribution is a comment organization framework:



Extraction of Implicit Aspects: Adjective-to-aspect Mapping

Idea from (Zhang & Liu, 2014)

"Tesco is large"

In other comments, we found as aspects of Tesco, qualified as "large":

- employer (2x)
- back office (1x)
- call center operation (1x)

We conclude: most probably, the employer aspect of Tesco was meant.

Extraction of Implicit Aspects: Context Words

Suppose now we need a **tiebreaker**!

rdf:type

"Tesco is large and the pay is good" **context words: {large, pay, company}**

In other comments, we found as aspects of Tesco, qualified as "large":

- employer (2x) **context words: {employer, large, salary, job, people}**
- back office (2x) **context words: {back, office, large, area, building}**
- call center operation (1x)

We use **context-words difference**, computed using **WordNet:Similarity**.

We conclude: most probably, the employer aspect of Tesco was meant.

Extraction of Implicit Aspects: Experiment and Result (1)

Adjective-to-aspect mapping			with context words	
Data	**Precision**	**Recall**	**Precision**	**Recall**
All News	76.87	26.12	88.65	30.03

Extraction of Implicit Aspects: Lexical Mapping

"Tesco is large"

In addition to "large", count also aspects qualified with similar adjectives ("huge", "big", …)

Similarity is measured using

1. lexical relationship <synonym> and <similar to> in **WordNet**
2. **WordNet::Similarity**

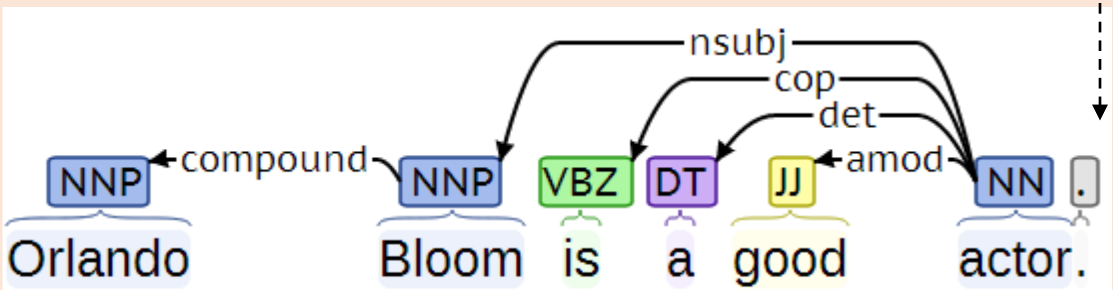
Extraction of Implicit Aspects: Experiment and Result (2)

Lexical mapping (1-step)			2-steps	
Data	**Precision**	**Recall**	**Precision**	**Recall**
All News	87.07	32.72	83.33	32.76

Extraction of Semi-Implicit Aspects (2)

- If we don't find the aspect within 1 step of lexical search:
- "Orlando Bloom is a good actor"

Stanford CoreNLP Annotation



Lexical database search

act — WordNet:derivationally related form — actor

act — WordNet:derivationally related form — acting

To find intermediate synsets, we consider following lexical relations:
synonym > antonym > similar to > derivationally related form (drf) > see also

We search for the closest noun to the pseudo-aspect.

Finding Aspects

1. Find words that represent the aspect

   job, beautiful, large, acted

   <noun>,<adjective>, or <verb>

Technique: grammar analysis
Tool: Stanford CoreNLP

2. Identify the aspect

   job, beauty, employer, acting

Technique: frequency-based mapping (implicit)
lexical database search (semi-implicit)
Tools: WordNet, DBpedia