CS 586 Final Project

Content Selection from Semantic Web Data

Presented by:

Sabita Acharya

12 December, 2014

New

Outline

- Project Summary
- Definitions
- Project Outline
- Starting Point
- Dataset
- Project Steps
- Results
- Comparison with related work
- Challenges faced
- Bibliography

Project Summary

- Based on the "Content selection from semantic web data" challenge[1].
- **Goal**: Build a system which given a set of RDF triples containing facts about celebrity and a target text, selects those triples that are reflected in target text [1].

Definitions

- **Content Selection**
  - Determines what subset of a large amount of information to include in the generated document.

Example:

- Automatically generating game summaries, given a database containing statistics on American Football[2].
- Generating recipes with respect to two different audiences: cooking novices and advanced cooks [3]

Definitions*(continued…)*

- **RDF triples**

<http://example.org/#spiderman>
<http://www.perceive.net/schemas/relationship/enemyOf>
<http://example.org/#green-goblin>

Linked Data Cloud comprises of more than 30 billion RDF triples.[2]

[2][Hasso Plattner Institute, "State of LOD Cloud", 2011, http://www4.wiwiss.fu-berlin.de/lodcloud/state/]

New

Definitions*(continued…)*

- **Part of Speech Tagging**

Input: Obama was born in US

Output: Obama/NNP, was/VBD, born/VBN, in/IN, US/NNP

- **Dependency Parsing(Collapsed)**

Project Outline

- **Given a set of DBpedia or Freebase triples and Wikipedia content as target text, we want to :**

- Find out different lexicalizations that appear for a particular property in the target text.

Example: "**dbpedia : spouse**" can be expressed by phrases like "**to be married to**" or "**to be the wife of"**, etc.

- Determine which specific line in the target text reflects the input triples.

Project Outline *(continued…)*

- George Walker Bush (born July 6, 1946) is an American politician and businessman who served as the 43rd President of the United States from 2001 to 2009, and the 46th Governor of Texas from 1995 to 2000. The eldest son of Barbara and George H. W. Bush, he was born in New Haven, Connecticut. After graduating from Yale University in 1968 and Harvard Business School in 1975, Bush worked in oil businesses.

dbpedia-owl:activeYearsEndDate	▪ 2000-12-21 (xsd:date) ▪ 2009-01-20 (xsd:date)
dbpedia-owl:activeYearsStartDate	▪ 1995-01-17 (xsd:date) ▪ 2001-01-20 (xsd:date)
dbpedia-owl:almaMater	▪ dbpedia:Yale_College ▪ dbpedia:Harvard_Business_School
dbpedia-owl:birthDate	▪ 1946-07-06 (xsd:date)
dbpedia-owl:birthPlace	▪ dbpedia:New_Haven,_Connecticut
dbpedia-owl:birthYear	▪ 1946-01-01 (xsd:date)
dbpedia-owl:bnfId	▪ 135678148
dbpedia-owl:child	▪ dbpedia:Jenna_Bush_Hager ▪ dbpedia:Barbara_Pierce_Bush
dbpedia-owl:individualisedGnd	▪ 12145391X
dbpedia-owl:lccnId	▪ no/95/49848
dbpedia-owl:lieutenant	▪ dbpedia:Rick_Perry ▪ dbpedia:Bob_Bullock
dbpedia-owl:militaryBranch	▪ dbpedia:Texas_Air_National_Guard ▪ dbpedia:Alabama_Air_National_Guard

Figure 1: A list of properties extracted from DBpedia

[Wikipedia, "About: George W. Bush", 2014, http://dbpedia.org/page/George_W._Bush]

Starting Point

- **The starting point for my project were the following papers :**

- Bouayad-Agha et al. explain about content selection challenge in detail[1].
- Barzilay et al. and Cimiano et al. use content selection to generate natural language texts[2,3].
- Walter et al. and Fabian et al. address a part of our problem and describe an approach that can be used to extract relations from texts[4,5] .

New

DataSet

- Used dataset provided by Content Selection Challenge.
- Total number of triples: **18307**
- Total number of sentences: **4988**
- Number of distinct triples : **613**
- Only **11** are present in over **40** percent of the files and only **19** predicates are present in over **10** percent of the files.
- Large number of predicates are present only in a few files.
- **40** percent of text files only contain one or two sentences.

New

DataSet

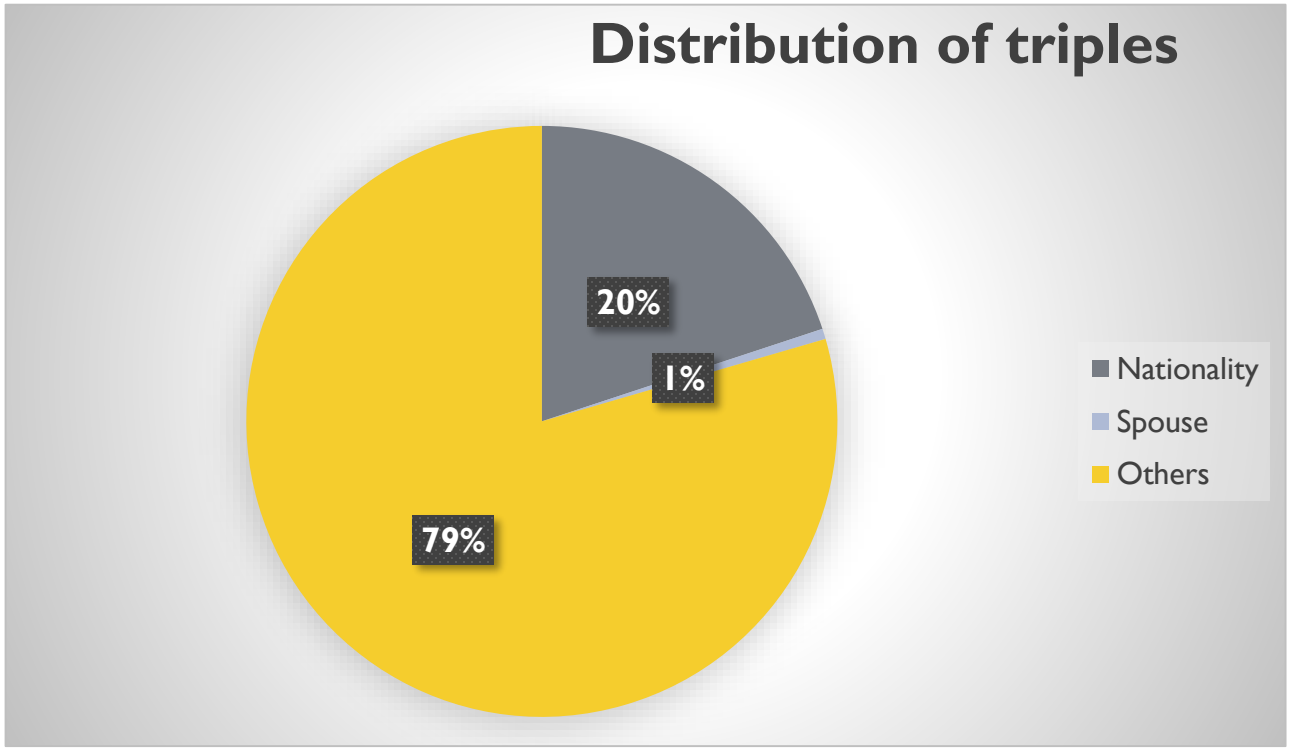

Figure 2: Distribution of triples in the challenge dataset

Data is very sparse

No proper formatting of triples.

Much time spent in extracting data

New

Project Steps

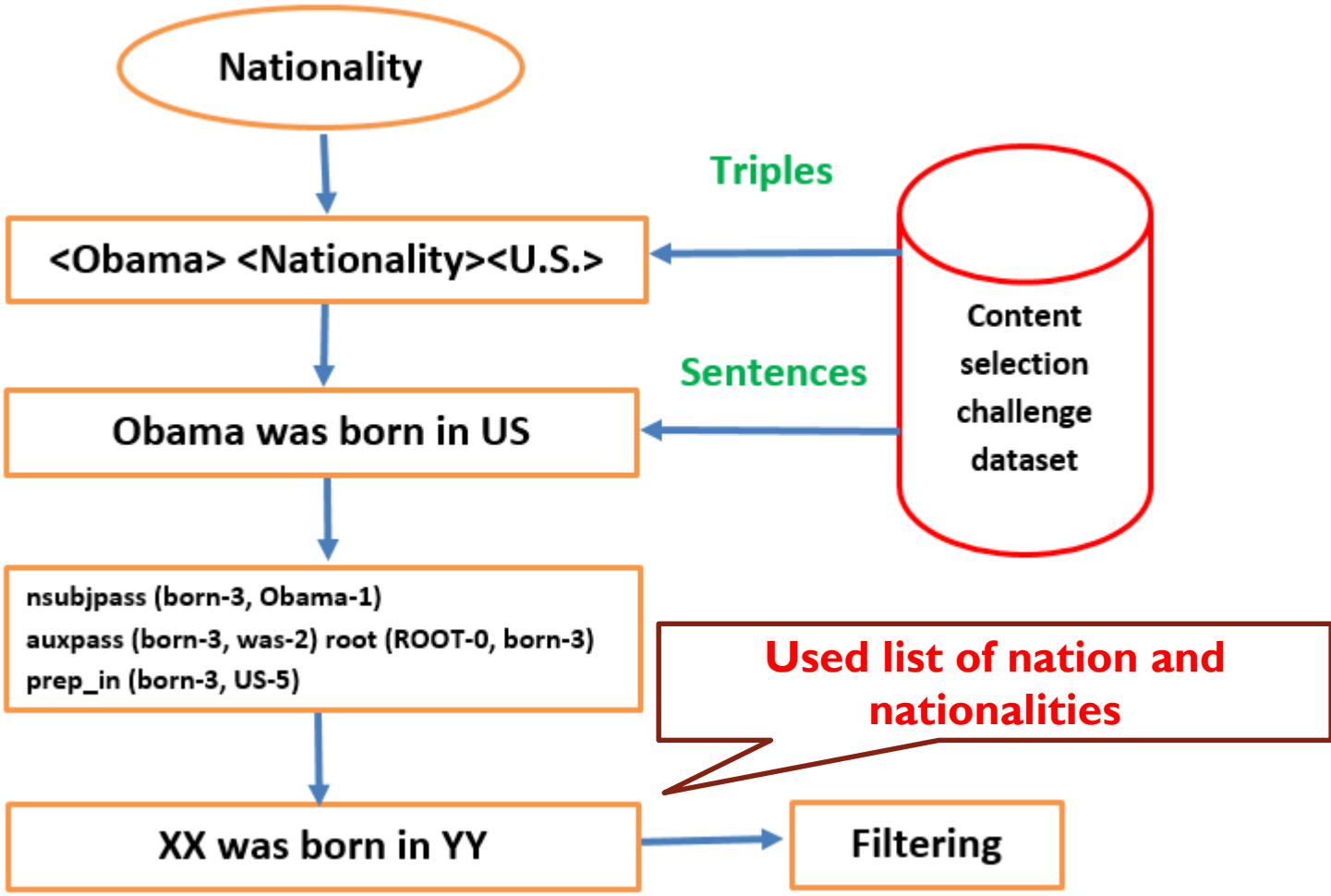

Figure 2: System Overview

Project Steps *(continued…)* New

❖ Filtering: extract features *(inspired from [5])*

XX was born in YY##Pos

[XX/NNP, was/VBD, born/VBN, in/IN, YY/NNP]

{(nsubjpass-right-VBN) } { (prep_in-left-VBN)}

({(nsubjpass-right-VBN) } { (prep_in-left-VBN)}{+1})

Project Steps *(continued…)* New

❖ Filtering: kNN algorithm

F1: ({(nsubjpass-right-VBN)} {(prep_in-left-VBN)})
F2: ({(dobj-left-NN)} {(advmod-right-RB)})

C1: {(nsubjpass-right-VBN)} O1: {(prep_in-left-VBN)}
C2: {(dobj-left-NN)} O2:{(advmod-right-RB)}

As mentioned in [5],

$$sim(C_1, C_2) = \sum_{\substack{(con_1, dir_1, w_1) \in C_1 \\ (con_2, dir_2, w_2) \in C_2}} \frac{\alpha_1(con_1 \sim con_2) + \alpha_2(dir_1 \sim dir_2) + \alpha_3 sim(w_1, w_2)}{|C_1| \cdot |C_2|}$$

con_i=context , dir_i=direction , w_i=POS tag

α_1=0.4, α_2=0.2, α_3=0.4

New

Project Steps *(continued…)*

❖ Filtering: kNN algorithm

F1: ({(nsubjpass-right-VBN)} {(prep_in-left-VBN)})

F2: ({(dobj-left-NN)} {(advmod-right-RB)})

sim(F1,F2)=(1/2)(sim(C1,C2)+sim(O1,O2))

- 10 fold cross validation to generate train and test data set.
- Compare similarity of a test feature with the entire positive and negative training features.
- Find out the sum of the top 10 similarity values.
- If similarity with positive training set is higher, assign positive else assign negative class.

New

Results

Property	Precision	Recall	fscore
Nationality	**81.26**	**76.36**	**78.14**
Spouse	71.49	87.14	77.98

Table 1: Results obtained by our system

Property	Precision	Recall	fscore
Nationality	79.83	51.25	62.40

Table 2: Results obtained by Venigalla et al. [6]

Comparison with Related Work

New

Systems	Input data	Parser used	Filtering techniques
Walter et al. [4]	Wikipedia, DBpedia	Malt Dependency Parser	-
Suchanek et al. [5]	-	Link Grammar Parser	kNN, SVM
Venigalla et al. [6]	Content selection challenge dataset	-	Cluster predicates, Use rules
Kutlak et al. [7]	Content selection challenge dataset	-	Use Google API
Our system	Content selection challenge dataset	Stanford Dependency Parser	kNN

Table 3: Comparison with other systems

Challenges faced

- Data was grouped according to the name of the person.
- Triples were provided in a .ttl format and had to be converted into .tsv format.
- Files for triples only contained predicates and objects. We extracted subjects from the names of the files.
- The object values for different predicates were given in a different format. So a common method of extraction would not work.
- Unavailability of sufficient amount of interesting data. Manually added some data so as to obtain a representative dataset.

Bibliography

[1] N. Bouayad-Agha, G. Casamayor, L. Wanner and C. Mellish, "Content selection from semantic web data, " *in Proceedings of International Natural Language Generation Conference,* 2012, pp. 146–149.

[2] R. Barzilay and M. Lapata, "Collective Content Selection for Concept-to-Text Generation," in *Proceedings of the Joint Human Language Technology and Empirical Methods in Natural Language Processing Conferences*, 2005.

[3] P. Cimiano, J. Luker, D. Nagel and C. Unger, "Exploiting ontology lexica for generating natural language texts from rdf data," in *Proceedings of the 14th European Workshop on Natural Language Generation*, 2013, pp. 10–19.

[4] S. Walter, C. Unger, and P. Cimiano, "A corpus-based approach for the induction of ontology lexica," in *Proceedings Of the 18th International Conference on Applications of Natural Language to Information Systems*, 2013, pp. 102–113.

Bibliography

[5] F.M. Suchanek, G. Ifrim, G. Weikum, "Combining linguistic and statistical analysis to extract relations from web documents," in *Proceedings of the 12th Association for Computing Machinery International conference on Knowledge discovery and data mining*, 2006.

[6] H. Venigalla, B.D. Eugenio, "UIC-CSC: The Content Selection Challenge Entry from the University of Illinois at Chicago," in *Proceedings of the 14th European Workshop on Natural Language Generation*, 2013, pp. 210–211.

[7] R. Kutlak, C. Mellish, K.V. Deemter, "Content Selection Challenge - University of Aberdeen Entry," in *Proceedings of the 14th European Workshop on Natural Language Generation*, 2013, pp. 208–209.