Tâche de English Slot Filling (Knowledge Base Population)

Dans le cadre de la campagne Text Analysis Conference

Kevin Lange Di Cesare
Département de génie informatique et génie logiciel
École Polytechnique de Montréal
Montréal, QC H3T 1J4
kevin.lange-di-cesare@polymtl.ca

Plan de la présentation

1. Introduction
2. TAC KBP English Slot Filling 2013
    1. Description de la tâche
    2. Requêtes
    3. Sortie des systèmes
    4. Évaluation
    5. Différentes approches
3. Système d'extraction de relations
    1. Description d'un système basé sur la supervision distante
    2. Composantes
    3. Survol des autres approches
4. Évolution de la tâche
5. Conclusion

Extraction d'information

- Extraction d'information (EI): repérér automatiquement des connaissances structurées sur un sujet d'intérêt à partir d'un corpus de textes

- Comprend deux phases:
  - Détection d'entités nommées
  - Détection des relations entre les entités nommées

Exemple:

With a father from **Kenya** and a mother from **Kansas**, **President Obama** was born in **Hawaii** on **August 4, 1961**

Entités nommées: Kenya, Kansas, President Obama, Hawaii, August 4, 1961

Relations: place_of_birth(President Obama, Hawaii),
date_of_birth(President Obama, August 4, 1961)

English Slot Filling

- Text Analysis Conference – Knowledge Base Population (TAC KBP)
  - English Slot Filling
- Objectif: *Promouvoir la recherche et évaluer la capacité d'un système automatisé à découvrir de l'information concernant des entités nommées et incorporer cet information à une source de connaissance.*
- Base de Connaissance: Base de données composée d'information structuée soit des triplets de la forme {Sujet, Précicat, Objet}
- Ressources:
  - Base de connaissance:
    - 2008 Wikipedia Snapshot (~800K entrées)
  - Corpus de documents:
    - Newswire: 1M
    - Web Documents : 1M
    - Forums: 100K

Example of a Wikipedia infobox entry : Barack Obama [subject]

The attribute *Born* can be declined to multiple relations:

[predicate]: [object]

- Name at birth : Barack Hussien Obama II
- Date of birth: August 4, 1961
- Age: 53
- City_of birth: Honolulu
- State of province of birth: Hawaii
- Country of birth: United-States

The relations are extracted from sentences in the associated Wikipedia article:

Barack Hussein Obama II *(US Listeni/bəˈrɑːk huːˈseɪn əˈbɑːmə/,* ***born August 4, 1961)*** *is the 44th and current President of the United States, and the first African American to hold the office.* ***Born*** *in* ***Honolulu****,* ***Hawaii****,* ***Obama*** *is a graduate of Columbia University and Harvard Law School, where he served as president of the Harvard Law Review.*

Relations (referred to as Slots)

- Total of 41 Wikipedia infoboxes slots
  - 21 slots for entity type PERSON
  - 16 slots for entity type ORGANISATION

Type	Slot Name	Content	Quantity
PER	per:alternate_names	Name	List
PER	per:children	Name	List
PER	per:cities_of_residence	Name	List
PER	per:city_of_birth	Name	Single
PER	per:city_of_death	Name	Single
PER	per:countries_of_residence	Name	List
PER	per:country_of_birth	Name	Single
PER	per:country_of_death	Name	Single
PER	per:employee_or_member_of	Name	List
PER	per:origin	Name	List
PER	per:other_family	Name	List
PER	per:parents	Name	List
PER	per:schools_attended	Name	List
PER	per:siblings	Name	List
PER	per:spouse	Name	List
PER	per:stateorprovince_of_birth	Name	Single
PER	per:stateorprovince_of_death	Name	Single
PER	per:statesorprovinces_of_residence	Name	List
PER	per:age	Value	Single
PER	per:date_of_birth	Value	Single
PER	per:date_of_death	Value	Single
PER	per:cause_of_death	String	Single
PER	per:charges	String	List
PER	per:religion	String	Single
PER	per:title	String	List

Type	Slot Name	Content	Quantity
ORG	org:alternate_names	Name	List
ORG	org:city_of_headquarters	Name	Single
ORG	org:country_of_headquarters	Name	Single
ORG	org:founded_by	Name	List
ORG	org:member_of	Name	List
ORG	org:members	Name	List
ORG	org:parents	Name	List
ORG	org:political_religious_affiliation	Name	List
ORG	org:shareholders	Name	List
ORG	org:stateorprovince_of_headquarters	Name	Single
ORG	org:subsidiaries	Name	List
ORG	org:top_members_employees	Name	List
ORG	org:date_dissolved	Value	Single
ORG	org:date_founded	Value	Single
ORG	org:number_of_employees_members	Value	Single
ORG	org:website	String	Single

Tableau 1 - Relations faisant l'objet de requête dans le cadre de la campagne TAC KBP English Slot Filling 2013

Relations (exemples)

Ramazan Bashardost (PER)

Age: 43

-> *A lawmaker and former Cabinet minister, **Bashardost**, **43**, is a self-styled populist and ascetic whose campaign office is a tent pitched outside the parliament building.*

Alternate names: Mohammad Ramzan Bashardost,

-> *A critic of government and parliamentarian **Mohammad Ramzan Bashardost**, with nearly 8,000 votes, secured the third position*

Cities of residence: Kabul, Paris

-> Ten percent chose independent candidate **Ramazan Bashardost**, a popular **Kabul** lawmaker who has campaigned to abolish corruption but is considered an eccentric, and six percent Ghani, said the US-based pro-democracy organisation

-> When the Taliban regime was ousted in 2001, **Bashardost** took a diplomatic post at the Afghan embassy in **Paris**, and in 2003 returned home to head the European Affairs Department at the Foreign Affairs Ministry.

Schools attended: Toloz University

-> **He** studied there and received his PhD in political science from **Toloz University**.

Requêtes

- 100 requêtes au total (50 personnes & 50 organisations)
  - Nom
  - Type
  - Document
  - Offsets de début et de fin
  - Node ID
  - Slots à ne pas remplir

Exemple:

```
<query id="SF_002">
   <name>PhillyInquirer</name>
   <docid>eng-NG-31-141808-9966244</docid>
   <beg>757</beg>
   <end>770</end>
   <enttype>ORG</enttype>
   <nodeid>E0312533</nodeid>
   <ignore>org:city_of_headquarters org:country_of_headquarters org:date_founded
org:number_of_employees_members org:stateorprovince_of_headquarters org:website</ignore>
</query>
```

Sortie des systèmes

Le système doit fournir au moins une réponse pour chaque pare Requête/Relation

Exemples:

Query: Ramazan Bashardost (PER) / Slot: Age

*A lawmaker and former Cabinet minister, **Bashardost**, **43**, is a self-styled populist and ascetic whose campaign office is a tent pitched outside the parliament building.*

Answer: 43

Query: Tahawwur Hussain Rana (PER) / Slot: Title

*Picture CHICAGO: Chicago **businessman Tahawwur Hussain Rana**, 49, is arraigned on charges he helped plot the deadly 2008 Mumbai attacks and an attack on a Danish newspaper in 2005.*

Answer : buisnessman

***Tahawwur Hussain Rana**, the **owner** of a Chicago-based immigration service, was already facing a possible 30-year sentence in connection with plans to attack a Danish newspaper whose cartoons were offensive to much of the Islamic world.*

Answer: owner

Évaluation

- Les réponses de tous les systèmes seront receuillies, annotées par des évaluateurs humains et répertoriées dans le gold standard.
- Le *slot-filler* pour chaque réponse sera évalué comme *Correct, Inexact, Redundant* ou *Wrong*

Correct = nombre total de classes d'équivalence correctes dans la réponse du système

System = nombre total de réponses non-NIL du systèeme

Reference = nombre de slots uni-valués pour lesquelles il y a une réponse non-NIL + nombre de classes d'équivalence pour chacun des slot multi-valués

Rappel = *Correct* / *Reference*

Précision = *Correct* / *System*

F-Measure = (2 * Précision * Rappel) / (Précision + Rappel)

Liste des équipes participantes

- 18 équipes ont soumis des résultats pour la tâche d'English Slot Filling

Team Id	Organization(s)	SF?	TSF?
ARPANI	Bhilai Institute of Technology	√	
CMUML	Carnegie Mellon University	√	√
PRIS2013	Beijing University of Posts and Telecommunications	√	
TALP_UPC	TALP Research Center of Technical University of Catalonia (UPC)	√	
UWashington	Department of Computer Science and Engineering, University of Washington	√	
utaustin	University of Texas at Austin – AI Lab	√	
SINDI	Korea Institute of Science and Technology Information	√	
CohenCMU	Carnegie Mellon University	√	
UMass_IESL	University of Massachusetts Amherst, Information Extraction and Synthesis Lab	√	
BIT	Beijing Institute of Technology	√	
SAFT_KRes	University of Southern California Information Sciences Institute	√	
UNED	Universidad Nacional de Educación a Distancia	√	√
IIRG	University College Dublin	√	
NYU	New York University	√	
Stanford	Stanford University	√	
lsv	Saarland University	√	
Compreno	ABBYY	√	√
RPI-BLENDER	Rensselaer Polytechnic Institute	√	√
MS_MLI	Microsoft Research		√

Résultats

- F1 médian: 15.7

	Diagnostic Scores			Official Scores		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
lsv	**32.93**	38.50	35.50	**33.17**	42.53	**37.28**
ARPANI*	29.10	47.83	**36.18**	27.45	50.38	35.54
RPI-BLENDER	30.62	38.19	33.98	29.02	40.73	33.89
PRIS2013	27.82	35.33	31.13	27.59	38.87	32.27
BIT	22.06	57.86	31.94	21.73	61.35	32.09
Stanford	28.46	32.30	30.26	28.41	35.86	31.70
NYU	17.35	50.70	25.85	16.76	53.83	25.56
UWashington	10.31	**59.72**	17.59	10.29	**63.45**	17.70
CMUML	10.63	28.79	15.53	10.69	32.30	16.07
SAFT_KRes	13.43	12.43	12.91	14.99	15.67	15.32
UMass_IESL	18.47	9.43	12.48	18.46	10.88	13.69
utaustin	7.91	21.85	11.62	8.11	25.16	12.26
UNED	9.11	15.08	11.36	9.33	17.59	12.19
Compreno	13.19	8.69	10.48	12.74	9.74	11.04
TALP_UPC	9.67	6.54	7.81	9.81	7.69	8.62
IIRG	3.20	7.38	4.46	2.86	7.72	4.17
SINDI	2.80	7.26	4.04	2.59	7.84	3.89
CohenCMU	3.68	1.69	2.32	3.68	1.98	2.57
LDC	58.35	83.81	68.80	57.08	85.60	68.49

Résultats de la tâche de Slot Filling pour les 100 entités dans l'ensemble de test

Comment ses résultats sont-ils obtenus?

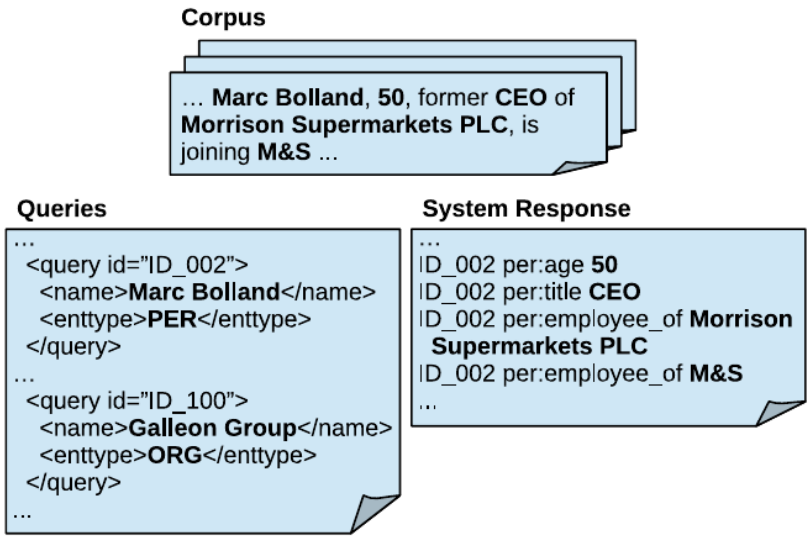
En utilisant un système d'extraction de relations:

- Le système *parse* le corpus de documents
- Pour une requête donnée, le système effectue une recherche parmi le corpus annoté et retourne les documents contant une référence à la requête
- Le système évalue si la phrase contenant l'entité-requête représente la relation voulue
- Le système retourne par la suite la réponse à la requête ainsi que la phrase dans laquelle elle a été trouvée comme justification

Système d'extraction de relations

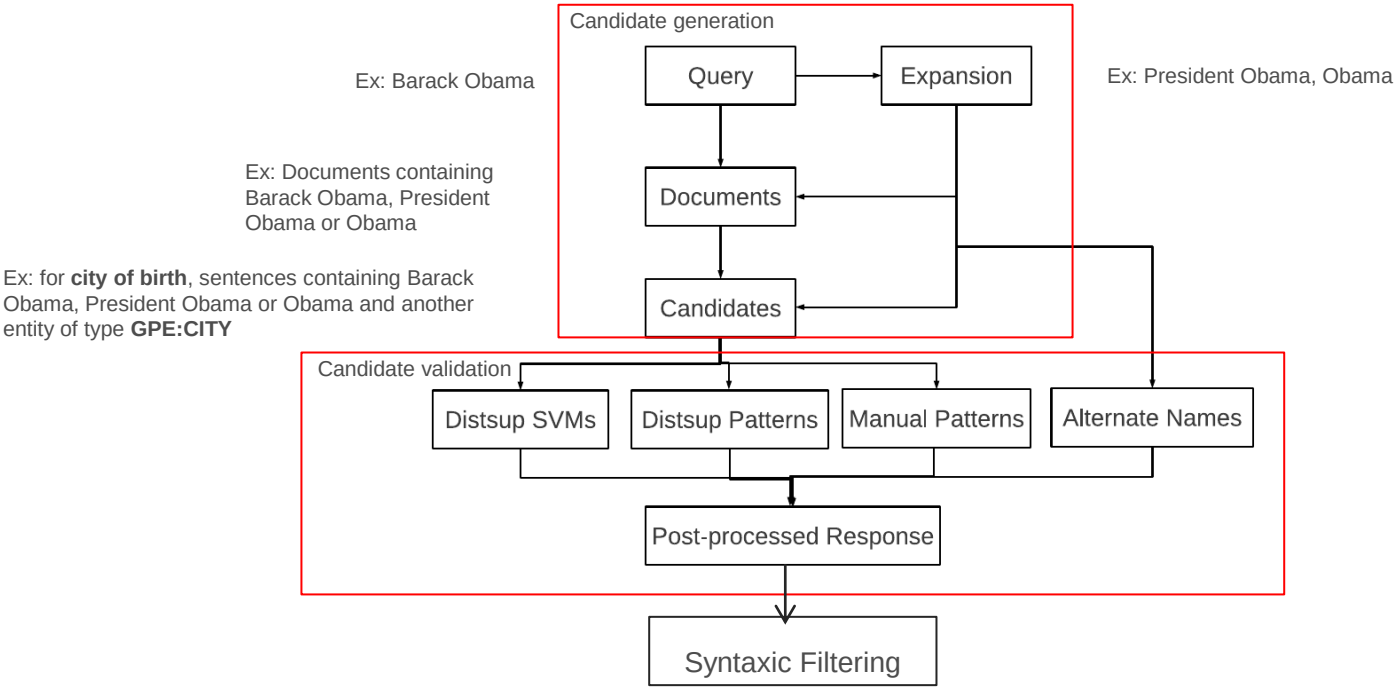
Relation Factory:

- Modulair
- Facilement configurable
- Open source
- Basé sur la supervison distante



Dataflow

- 2 phases:
  - Candidate generation: document retrieval, sentence filtering based on query match
  - Candidate validation



Génération des candidats

- Les documents sont récupérés en utilisant la requête originale ainsi que son expansion
- Les phrases sont annotées utilisant un *named-entity tagger (state-of-the sequence perceptron named-entity tagger [Chrupala and Klakow, 2010]).* Les types non-standard d'entité nommée sont appariés en utilisant une liste de types provenant de Freebase
- Les phrases pour lesquelles une requête ou une de ses expansions correspondent et contiennent une entité nommée du type d'un potientiel *slot filler* sont passées à la phase de validation

Exemples d'expansion :

Original query	Wikipedia link anchor text expansions	Per: last name / Org: suffixes
Ali Akbar Khan	Utd. Ali Akbar Khan, Ustad Ali Akbar Khan	Khan
Adam Gadahn	Azzam the American, Adam Yahiye Gadahn	Gadahn
Nancy Kissel	Murder of Robert Kissel, Robert Kissel	Kissel
DCNS	Direction des Constructions Navales, DCN, ...	DCNS Ltd, DCNS Corp, ...
STX Finland	Kvaerner Masa Yards, Aker Finnyards, ...	STX Finland Ltd, ...

Exemple

Query : Ramazan Bashardhost (PER)

Query Expansions: Bashardost

Slot: Cities of residence

Expected filler tag : **GPE:CITY**

Sentence : Ten percent chose independent candidate **Ramazan Bashardost**, a popular **Kabul** lawmaker…

Ten	B-CARDINAL
percent	O (other)
chose	O
independent	O
candidate	O
Ramazan	**B-PERSON**
Bashardost	**I-PERSON**
a	O
Popular	O
Kabul	**B-GPE:CITY**
Lawmaker	O

The sentence is retained and sent to validation stage because it contains the query or the query expansion and a word with the exepected filler tag (GPE:CITY in this case)

Distant Supervision

Supervised learning for which examples (training data) are not explicitly available

Example:

Relation: date of birth

Idea: Use pairs of arguments within a **knowledge base** which are in relation by date of birth.

- Barack Obama (ARG1): August 4, 1961 (ARG2)
- Ramazan Bashardost (ARG1): 1965 (ARG2)
- John Lennon (ARG1): October 9, 1940 (ARG2)
- Wayne Gretzky (ARG1): January 26, 1961 (ARG2)

For each pair arguments, extract sentences from a **text corpus** in which both arguments are present:

Pair of arguments: {Barack Obama (ARG1), August 4, 1961 (ARG2)}

Sentences extracted:

- **Barack Obama** was born in Hawaii on **August 4, 1961**.
- **Barack Obama**'s birth date is **August 4, 1961**.
- Born on **August 4, 1961 Barack Obama** is the 44th president of the United States.

These sentences are retained as positive training examples from which various features can be extracted for classification

Classifieurs SVM basés sur la supervision distante

- Données d'entraînement:
  - Les paires d'arguments pour la supervision distante sont obtenues en mappant des relations de Freebase aux relations de TAC KBP et en appariant des modèles générés manuellement text collections. (max 10k phrases par relation)
  - Max 10k paires d'arguments par relations, et max 500 phrases par paire d'arguments
  - Pour une relation donnée, les phrases contenant les deux arguments sont retenues comme exemples positifs alors que les exemples pour les autres relations sont utilisés comme exemples négatifs

- Features:
  - Précédence de la réponse par la requête
  - Token n-grams (3)
  - Skip n-grams (3-4)
  - Contexte avant, entre et après les entités
- 1 SVM binaire par relation (SVMlight)

Modèles générés manuellement

- Utilise des modèles provenant de définitions et d'exemples se trouvant directement dans la description de la tâche
- Les modèles sont également utilisés pour extraire des données d'entraînement supplémentaire pour la supervision distante

Exemple:

Relation: per:stateorprovince_of_birth

Phrase: Harper, born in April of 1959 in Toronto, Ontario

Modèle: *ARG1, born * in *, ARG2*

ARG1: entité-requête

ARG2: entité-réponse

(*): 1 à 4 tokens

Différentes approches:

- Supervision distante et règles:
  - Implémenté de manière séparée et combinant les sorties (NYU, Standford, BIT)
  - En utilisant des règles afin de générer des données d'entraînement supplémentaires pour la supervision distante (lsv)
- Modèles de supervision distante plus complexes avec réduction de bruit intégré (Stanford)
- *Open-domain information extraction* utilisant Open IE 4.0(UWashington)
  - Extraction de tuples de format (Arg1, Rel, Arg2) du corpus de KBP. Utilisation de règles pour mapper ces tuples aux relations spécifiques à la tâche
- *Bootstrapping* de modèles basés sur des chemins d'arbres de dépendance utilisant des tuples d'entités-requête et entités-réponses de la base de connaissance comme *seeds* (PRIS2013)
- Apprentissage non-supervisé

TAC KBP 2014

Changements:

- Intégration de requêtes ambigües
- Réponses requiérant de l'inférence (inter-document)
- Aucun lien vers la base de connaissance. Aucune relation à ignorer

Résultats:

- Performances similaires à l'année précédente (F1 médian 2014: 19.8 2013: 15.71)
- Meilleures performances: Stanford:
  - Rappel: 27.66 %
  - Précision: 54.61 %
  - F1: 36.72 %

Conclusion

- Les systèmes s'améliorent:
  - La tâche de 2014 est plus complèxe qu'en 2013
  - Ex: Pour Stanford, la différence est considérable puisque leur performance augmente de 27.6 F1 en utiliant leur système de 2013 à 36.7 F1 en utilisant celui de 2014
- L'approche favorisée pour la tâche demeure la supervision distante

- Projet : Développer un module de filtrage syntaxique s'insérant en aval d'un système permettant d'améliorer les performances de celui-ci
  - améliorer les performances de *Relation Factory*, le système ayant obtenue les meilleures performances à la campagne TAC KBP English Slot Filling 2013

Références

[1] Benjamin Roth & al., «Effective Slot Filling Based on Shallow Distant Supervision Methods»,*Sixth Text Analysis Conference (TAC 2013)*, Gaitherburg, 2013.

[2] T. Heath et C. Bizer, «Linked Data: Evolving the Web into a Global Data Space,» chez *Synthesis lectures on the semantic web: theory and technology*, 2011, pp. 1-136.

[3] M. Surdeanu et H. Ji, Overview of the English Slot Filling Track at the TAC2014 Knowledge Base Population Evaluation

[4] M. Surdeanu, Overview of the TAC2013 Knowledge Base Population Evaluation: English Slot Filling and Temporal Slot Filling

[5] Benjamin Roth & al., «Generalizing from Freebase and Patterns using Cluster-Based Distant Supervision for KBP Slot-Filling*», Fifth Text Analysis Conference (TAC 2012)*