Tâches à saveur sémantiques

felipe@iro.umontreal.ca

RALI
Dept. Informatique et Recherche Opérationnelle
Université de Montréal

V0.5

Last compiled: 22 novembre 2018
Plan

Sémantique

Question-réponse
  ARC
  RACE
  WDW
Plan

Sémantique

Question-réponse
  ARC
  RACE
  WDW
la plupart de ces ressources (et d’autres) sont listées ici :
http://www.cs.cmu.edu/~mfaruqui/suite.html
Rubenstein and Goodenough

http://www.cs.cmu.edu/~mfaruqui/word-sim/EN-RG-65.txt

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>gem</td>
<td>jewel</td>
</tr>
<tr>
<td>2</td>
<td>midday</td>
<td>noon</td>
</tr>
<tr>
<td>3</td>
<td>automobile</td>
<td>car</td>
</tr>
<tr>
<td>4</td>
<td>cemetery</td>
<td>graveyard</td>
</tr>
<tr>
<td>5</td>
<td>cushion</td>
<td>pillow</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>autograph</td>
<td>shore</td>
</tr>
<tr>
<td>8</td>
<td>fruit</td>
<td>furnace</td>
</tr>
<tr>
<td>9</td>
<td>noon</td>
<td>string</td>
</tr>
<tr>
<td>10</td>
<td>rooster</td>
<td>voyage</td>
</tr>
<tr>
<td>11</td>
<td>cord</td>
<td>smile</td>
</tr>
</tbody>
</table>

felipe@iro.umontreal.ca  
Tâches à saveur sémantiques
WordSimilarity-353

  - **set1** 153 paires de mots : 13 jugements de relatedness
  - **set2** 200 paires de mots : 16 jugements

<table>
<thead>
<tr>
<th>1</th>
<th>% more set1.csv</th>
<th>sort -t, -k3,3n+</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>tiger,tiger,10.00,10,10,10,10,10,10,10,10,10,10,10,10,10</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>coast,shore,9.10,9,9.75,10,10,6,6,10,8,9.5,10,10,10,10</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>football,soccer,9.03,9,9.9,9,10,8,7,9.5,10,8,10,9,9,9</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>fuck,sex,9.44,10,9.75,10,10,8,9,9,10,9,10,8,10,10</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>journey,voyage,9.29,9,9.75,10,10,6,7,9.5,10,9.5,10,10,10</td>
<td></td>
</tr>
</tbody>
</table>

- **split de** [http://alfonseca.org/eng/research/wordsim353.html](http://alfonseca.org/eng/research/wordsim353.html)

  - similarité 203 paires
  - relatedness 252 paires
À propos du split de [Agirre et al., 2009]

- 103 paires en commun dans les deux fichiers :

```bash
more wordsim_relatedness_goldstandard.txt \ 
wordsim_similarity_goldstandard.txt | \ 
sort | uniq -c | awk '$1 == 2 {print $0}' | \ 
sort -k4,4gr | head -n 5+
```

1. 2 cup  food  5,00  
2. 2 monk  oracle  5,00  
3. 2 journal  association  4,97  
4. 2 car  flight  4,94  
5. 2 street  children  4,94
MEN

http://clic.cimec.unitn.it/~elia.bruni/MEN.html

- 3000 paires de mots avec un score (entre 0 et 50) indiquant leur relatedness

<table>
<thead>
<tr>
<th>cat-n</th>
<th>feline-j</th>
<th>48</th>
<th>gun-n</th>
<th>stair-n</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>copper-n</td>
<td>metal-n</td>
<td>48</td>
<td>military-j</td>
<td>tomato-n</td>
<td>2</td>
</tr>
<tr>
<td>flower-n</td>
<td>garden-n</td>
<td>48</td>
<td>angel-n</td>
<td>gasoline-n</td>
<td>1</td>
</tr>
<tr>
<td>grass-n</td>
<td>lawn-n</td>
<td>48</td>
<td>feather-n</td>
<td>truck-n</td>
<td>1</td>
</tr>
<tr>
<td>pregnancy-n</td>
<td>pregnant-j</td>
<td>48</td>
<td>festival-n</td>
<td>whisker-n</td>
<td>1</td>
</tr>
</tbody>
</table>

- paires vérifiant des contraintes de fréquences dans différentes ressources et échantillonnées de manière à générer une tâche balancée

- évaluation faite par comparaison binaire à 50 autres paires sur Mecanical Turk

- tâche : trouver les paires les plus similaires (ranking)
TOEFL Tests [Freitag et al., 2005]

- accessible ici :
  http://www.cs.cmu.edu/~dayne/wbst-nanews.tar.gz

- chaque instance de test est de la forme :

  \[
  \text{stimulus} \quad \begin{cases}
    \text{target} : [\text{synonym and 3 distractors}] \\
    \text{ex :}
    \text{comprehend : [encompass career admire pardon]}
  \end{cases}
  \]

- 9887 instances de test pour les noms, 7398 pour les verbes
<table>
<thead>
<tr>
<th>Sem</th>
<th>QA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TOEFL Tests</strong> [Freitag et al., 2005]</td>
<td></td>
</tr>
<tr>
<td>force</td>
<td>violence</td>
</tr>
<tr>
<td>force</td>
<td>power</td>
</tr>
<tr>
<td>force</td>
<td>personnel</td>
</tr>
<tr>
<td>force</td>
<td>strength</td>
</tr>
<tr>
<td>force</td>
<td>effect</td>
</tr>
<tr>
<td>buy</td>
<td>bribe</td>
</tr>
<tr>
<td>buy</td>
<td>corrupt</td>
</tr>
<tr>
<td>buy</td>
<td>purchase</td>
</tr>
<tr>
<td>record</td>
<td>commemorate</td>
</tr>
<tr>
<td>record</td>
<td>enter</td>
</tr>
<tr>
<td>record</td>
<td>tape</td>
</tr>
</tbody>
</table>
Word2Vec datasets (analogies)

http://word2vec.googlecode.com/svn/trunk

questions-words.txt 19544 équations (avec solution) :

: capital-common-countries
  Berlin Germany Islamabad Pakistan
  Baghdad Iraq Oslo Norway
: capital-world
  Algiers Algeria Belmopan Belize
  Abuja Nigeria Doha Qatar
: currency
  Armenia dram India rupee
  Angola kwanza Denmark krone
: city-in-state
  Philadelphia Pennsylvania Atlanta Georgia
  Chicago Illinois Stockton California
: family
  father mother husband wife
  brother sister prince princess
: gram1-adjective-to-adverb
  cheerful cheerfully efficient efficiently
  apparent apparently free freely
: gram2-opposite
  clear unclear impressive unimpressive
  aware unaware efficient inefficient
: gram3-comparative
  bright brighter simple simpler
  big bigger bright brighter
: gram4-superlative
  bright brightest weird weirdest
  big biggest dark darkest
: gram5-present-participle
  debug debugging code coding
  dance dancing enhance enhancing
: gram6-nationality-adjective
  Australia Australian India Indian
  Albania Albanian Sweden Swedish
: gram7-past-tense
  describing described playing played
  dancing danced vanishing vanished
: gram8-plural
  bottle bottles man men
  bird birds bottle bottles
: gram9-plural-verbs
  enhance enhances play plays
  describe describes increase increases

felipe@iro.umontreal.ca

Tâches à saveur sémantiques
Word2Vec datasets (analogies)

http://word2vec.googlecode.com/svn/trunk

questions-phrases.txt  3218 équations (avec solution) :

: newspapers
    Boston  Boston_Globe  Salt_Lake  Salt_Lake_Tribune
    Baltimore  Baltimore_Sun  Cincinnati  Cincinnati_Enquirer

: ice_hockey
    Calgary  Calgary_Flames  Ottawa  Ottawa_Senators
    Boston  Boston_Bruins  Chicago  Chicago_Blackhawks

: basketball
    Dallas  Dallas_Mavericks  Boston  Boston_Celtics
    Boston  Boston_Celtics  New_York  New_York_Knicks

: airlines
    Finland  Finnair  Spain  Spanair
    Austria  Austrian_Airlines  Turkey  Turkish_Airlines

: people-companies
    Samuel_J._Palmisano  IBM  Jeff_Bezos  Amazon
    Eric_Schmidt  Google  Larry_Ellison  Oracle
À propos des datasets Word2Vec

http://word2vec.googlecode.com/svn/trunk

- 5031 analogies impliquant la relation capitale_of, mais seulement 118 paires de pays / capitale . . .
- 870 analogies impliquant la relation plural-verbs, mais seulement 30 paires de verbes . . .
  - 27 avec la relation +s ; ex : eat / eats
  - et seulement 3 avec la relation +es : go / goes, search / searches et vanish / vanishes
  - note : [go: goes :: eat : eats] n’est pas une analogie formelle
- 1560 analogies impliquant la relation past-tense, mais seulement 40 paires de verbes . . .
  - ex : decreasing / decreased, knowing / knew
  - dont la moitié (23) sont des verbes irréguliers ; ex : sitted / sat et slept

felipe@iro.umontreal.ca
MSR Sentence Completion dataset


▶ 1040 questions comme celles-ci :

26 It may be that the solution of the one may _____ to be the solution of the other.

   a) prove  
   b) learn  
   c) choose  
   d) forget  
   e) succumb

27 We shall just be in time to have a little _____ with him.

   a) breakfast  
   b) elegance  
   c) garment  
   d) basket  
   e) dog
MSR Sentence Completion dataset


- 1040 questions comme celles-ci :

26 It may be that the solution of the one may prove to be the solution of the other.

   a) prove  
b) learn  
c) choose  
d) forget  
e) succumb

27 We shall just be in time to have a little breakfast with him.

   a) breakfast  
b) elegance  
c) garment  
d) basket  
e) dog
BLESS

https://sites.google.com/site/geometricalmodels/shared-evaluation

- 26 554 lignes comme celle-ci :
  frog-n amphibian_reptile attri green-j
  qui indique qu’une grenouille est un amphibien ou un reptile et
  qu’un de ses attributs est d’être verte (gasp).

- 200 concepts (reptiles, mammifères, oiseaux, fruits, insectes,
  armes, etc.), 5 relations (attribut, co-hyponyme, hyperonyme,
  meronyme, événement)

  On average, there are 14 attri, 18 coord, 19 event,
  7 hyper and 15 mero relata per concept.

- 14 400 concept-relation-relatum + 12 154 control tuples (random)

une tâche suggérée
BLESS/frog

https://sites.google.com/site/geometricalmodels/shared-evaluation

<table>
<thead>
<tr>
<th>mero</th>
<th>blood-n bone-n eye-n foot-n gill-n head-n leg-n lung-n poison-n skin-n tongue-n wart-n</th>
</tr>
</thead>
<tbody>
<tr>
<td>event</td>
<td>breathe-v catch-v chase-v croak-v die-v drink-v eat-v hop-v kill-v lay-v leap-v live-v sit-v swim-v</td>
</tr>
<tr>
<td>attri</td>
<td>amphibious-j aquatic-j brown-j colorful-j edible-j green-j little-j loud-j noisy-j old-j poisonous-j small-j ugly-j wild-j</td>
</tr>
<tr>
<td>coord</td>
<td>alligator-n bullfrog-n lizard-n snake-n toad-n turtle-n</td>
</tr>
<tr>
<td>hyper</td>
<td>amphibian-n animal-n beast-n chordate-n creature-n vertebrate-n</td>
</tr>
</tbody>
</table>

+ 50 relations aléatoires

frog-n amphibian_reptile random-n doyen-n

felipe@iro.umontreal.ca Tâches à saveur sémantiques
SemEval 2010

Cause-Effect (1003)
He had chest pains and <el>headaches</el> from <e2>mold</e2> in the bedrooms.

Component-Whole (941)
The <el>provinces</el> are divided into <e2>counties</e2> (shahrestan), and ...

Entity-Destination (845)
The prosecutors carried <el>guns</el> into the <e2>court room</e2>.

Product-Producer (717)
Next the <el>soldiers</el> put up a <e2>wall</e2> of stakes on the pile of dirt.

+ Entity-Origin (716), Member-Collection (690), Message-Topic (634), Content-Container (540), Instrument-Agency (504)

Tâche : identifier la relation de 2717 paires de noms marquées dans des phrases
<table>
<thead>
<tr>
<th>MT</th>
<th>When we are faced with a potential risk, it is important to apply the precautionary principle.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>When we are faced with a potential risk, it is important to put into practice the precautionary principle.</td>
</tr>
<tr>
<td>MT</td>
<td>This rule can be applied equally to all recreational but dangerous drugs when no one but the person consuming the drug is likely to be affected.</td>
</tr>
<tr>
<td>4.8</td>
<td>This rule can, moreover, apply to all drugs récréationnelles but dangerous provided that the consumer of drugs is the only run the risks.</td>
</tr>
<tr>
<td>MSR-VID</td>
<td>A man stares out a window.</td>
</tr>
<tr>
<td>4.8</td>
<td>A man looks out the window.</td>
</tr>
<tr>
<td>MSR-PAR</td>
<td>But at age 15, she had reached 260 pounds and a difficult decision : It was time to try surgery.</td>
</tr>
<tr>
<td>4.2</td>
<td>But at the age of 15, she weighed a whopping 117kg and came to a difficult decision : it was time to try surgery.</td>
</tr>
</tbody>
</table>
SemEval 2014 / Task3

- **Phrase2Word** (*TRAIN*: 500, *TEST*: 500) — échelle \([0, 4]\)

<table>
<thead>
<tr>
<th>Score</th>
<th>Phrase</th>
<th>Lexical Unit</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>the structural and functional unit of all known living organisms</td>
<td>DNA</td>
<td>Lexicographic</td>
</tr>
<tr>
<td>4</td>
<td>traditionally the land on which a college or university and related institutional buildings are situate</td>
<td>campus</td>
<td>Lexicographic</td>
</tr>
<tr>
<td>4</td>
<td>watering hole</td>
<td>pub</td>
<td>Idiomatic</td>
</tr>
<tr>
<td>4</td>
<td>wear out one’s welcome</td>
<td>exhaust</td>
<td>Idiomatic</td>
</tr>
<tr>
<td>4</td>
<td>whole ball of wax</td>
<td>total</td>
<td>Idiomatic</td>
</tr>
<tr>
<td>0</td>
<td>to make from scratch</td>
<td>breath</td>
<td>Idiomatic</td>
</tr>
<tr>
<td>0</td>
<td>what a vivid imagination</td>
<td>insipid</td>
<td>Newswire</td>
</tr>
<tr>
<td>0</td>
<td>workouts to strengthen abdominal muscles</td>
<td>fractions</td>
<td>Search</td>
</tr>
</tbody>
</table>

felipe@iro.umontreal.ca  
Tâches à saveur sémantiques
SemEval 2014 / Task3

▶ Sentence2Phrase (TRAIN : 500, TEST : 500) — échelle [0, 4]

<table>
<thead>
<tr>
<th>Score</th>
<th>Sentence</th>
<th>Interpretation</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>We completely agree and there’s no daylight between us on the issue.</td>
<td>complete agreement</td>
<td>Idiomatic</td>
</tr>
<tr>
<td>4</td>
<td>do you now where i can watch free older movies online without download?</td>
<td>streaming vintage movies for free</td>
<td>CQA</td>
</tr>
</tbody>
</table>

▶ Paragraph2Sentence (TRAIN : 500, TEST : 500) — échelle [0, 4]

<table>
<thead>
<tr>
<th>Score</th>
<th>Paragraph</th>
<th>Question</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>so i bought an item a couple weeks ago on eBay and i still havent even been notified that my item has been shipped yet.. what do i do ?</td>
<td>What should I do when my eBay purchase still hasn’t shipped after two weeks ?</td>
<td>CQA</td>
</tr>
</tbody>
</table>
Plan

Sémantique

Question-réponse
  ARC
  RACE
  WDW
Quelques jeux de test pour le QR (QA)

- SQUAD
- MS Marco
- NewsQA
- MC Test
- Cloze-style
- RACE
- CNN/Daily
- WDW (What did what)
- AI2 Elementary school Science questions
- triviaQA
- WikiHop
ARC (AI2 Reasoning Challenge)  
[Clark et al., 2018]

- 7787 questions à choix multiples en sciences naturelles :
  Easy set : 5197 questions  
  Challenge set : 2590 questions qui résistent à des baselines basés sur la RI ou la co-occurrence

- questions pour étudiants de 8 à 13 ans

1. Which property of a mineral can be determined just by looking at it?  
   (A) luster (B) mass (C) weight (D) hardness

**note** : il est dit qu’il n’y a pas sur le web de phrase comme luster can be determined by looking at
What is saying Google about this?

Google search results for 'luster can be determined by looking at'

About 2,370,000 results (0.53 seconds)

What is Luster? - Definition, Types & Examples - Video & Lesson ...
https://study.com/academy/lesson/what-is-luster-definition-types-examples.html
It is one of the properties mineralogists look at when trying to determine the identity ... The greasy type of luster can be found in minerals that look like they were ...

People also ask

- How do you find the luster of a rock?
- What is the luster of metals?
- How do you test for Luster?
- What is shiny luster called?
ARC (AI2 Reasoning Challenge)  
*Clark et al., 2018*

- 7787 questions à choix multiples en sciences naturelles :  
  - Easy set : 5197 questions  
  - Challenge set : 2590 questions qui résistent à des baselines basés sur la RI ou la co-occurrence

- questions pour étudiants de 8 à 13 ans

1. Which property of a mineral can be determined just by looking at it?  
   (A) luster (B) mass (C) weight (D) hardness

2. A student riding a bicycle observes that it moves faster on a smooth road than on a rough road. This happens because the smooth road has  
   (A) less gravity (B) more gravity (C) less friction (D) more friction
ARC distribué avec

- 14M de phrases reliées (couvre 95% des connaissances requises pour les questions)
- 3 implémentations de réseaux profonds (top performers sur d’autres tâches de QA)
- http://data.allenai.org/arc
## Performances

<table>
<thead>
<tr>
<th>Solver</th>
<th>Test Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Challenge Set</td>
</tr>
<tr>
<td>IR (dataset defn)</td>
<td>(1.02)$\dagger$</td>
</tr>
<tr>
<td>PMI (dataset defn)</td>
<td>(2.03)$\dagger$</td>
</tr>
<tr>
<td>IR (using ARC Corpus)</td>
<td>20.26</td>
</tr>
<tr>
<td>TupleInference</td>
<td>23.83</td>
</tr>
<tr>
<td>DecompAttn$\ddagger$</td>
<td>24.34</td>
</tr>
<tr>
<td>Guess-all (“random”)</td>
<td>25.02</td>
</tr>
<tr>
<td>DGEM-OpenIE$\ddagger$</td>
<td>26.41</td>
</tr>
<tr>
<td>BiDAF$\ddagger$</td>
<td>26.54</td>
</tr>
<tr>
<td>TableILP</td>
<td>26.97</td>
</tr>
<tr>
<td>DGEM</td>
<td>27.11</td>
</tr>
</tbody>
</table>

$\dagger$ These solvers were used to define the dataset, affecting scores.

$\ddagger$ Code available at https://github.com/allenai/arc-solvers

Table 6: Performance of the different baseline systems. Scores are reported as percentages on the test sets. For up-to-date results, see the ARC leaderboard at http://data.allenai.org/arc.
RACE [Lai et al., 2017]

- 28k passages, 100k questions posées par des enseignants de l’anglais langue seconde à des étudiants Chinois agés de 12 à 18 ans
- deux niveaux : RACE-M (12-15 ans), RACE-H (15-18)
- vocabulaire restreint (apprenants de l’anglais)

- code : [https://github.com/qizhex/RACE_AR_baselines](https://github.com/qizhex/RACE_AR_baselines)

- voir aussi : MCTest [Richardson et al., 2013] (histoires fictives)
Passage:
In a small village in England about 150 years ago, a mail coach was standing on the street. It didn’t come to that village often. People had to pay a lot to get a letter. The person who sent the letter didn’t have to pay the postage, while the receiver had to.
“Here’s a letter for Miss Alice Brown,” said the mailman.
“I’m Alice Brown,” a girl of about 18 said in a low voice.
Alice looked at the envelope for a minute, and then handed it back to the mailman.
“I’m sorry I can’t take it, I don’t have enough money to pay it”, she said.
A gentleman standing around were very sorry for her. Then he came up and paid the postage for her.
When the gentleman gave the letter to her, she said with a smile, “Thank you very much, This letter is from Tom. I’m going to marry him. He went to London to look for work. I’ve waited a long time for this letter, but now I don’t need it, there is nothing in it.”
“Really? How do you know that?” the gentleman said in surprise.
“He told me that he would put some signs on the envelope. Look, sir, this cross in the corner means that he is well and this circle means he has found work. That’s good news.”
The gentleman was Sir Rowland Hill. He didn’t forget Alice and her letter.
“The postage to be paid by the receiver has to be changed,” he said to himself and had a good plan.
“The postage has to be much lower, what about a penny? And the person who sends the letter pays the postage. He has to buy a stamp and put it on the envelope.” he said. The government accepted his plan. Then the first stamp was put out in 1840. It was called the “Penny Black”. It had a picture of the Queen on it.

Questions:

1): The first postage stamp was made ...
A. in England B. in America C. by Alice D. in 1910

2): The girl handed the letter back to the mailman because ...
A. she didn’t know whose letter it was
B. she had no money to pay the postage
C. she received the letter but she didn’t want to open it
D. she had already known what was written in the letter

3): We can know from Alice’s words that ...
A. Tom had told her what the signs meant before leaving
B. Alice was clever and could guess the meaning of the signs
C. Alice had put the signs on the envelope herself
D. Tom had put the signs as Alice had told him to

4): The idea of using stamps was thought of by ...
A. the government
B. Sir Rowland Hill
C. Alice Brown
D. Tom

5): From the passage we know the high postage made ...
A. people never send each other letters
B. lovers almost lose every touch with each other
C. people try their best to avoid paying it
D. receivers refuse to pay the coming letters

Answer: ADABC
Performances

<table>
<thead>
<tr>
<th></th>
<th>RACE-M</th>
<th>RACE-H</th>
<th>RACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>24.6</td>
<td>25.0</td>
<td>24.9</td>
</tr>
<tr>
<td>Sliding Window</td>
<td>37.3</td>
<td>30.4</td>
<td>32.2</td>
</tr>
<tr>
<td>Stanford AR</td>
<td>44.2</td>
<td>43.0</td>
<td>43.3</td>
</tr>
<tr>
<td>GA</td>
<td>43.7</td>
<td>44.2</td>
<td>44.1</td>
</tr>
<tr>
<td>Turkers</td>
<td>85.1</td>
<td>69.4</td>
<td>73.3</td>
</tr>
<tr>
<td>Ceiling Performance</td>
<td>95.4</td>
<td>94.2</td>
<td>94.5</td>
</tr>
</tbody>
</table>

Table 5: Accuracy of models and human on the each d previous publications. DM denotes Daily Mail and WD

**SW** $\text{tf-idf de } \text{conc}(q, a) \text{ et d'un span candidat}$

**AR** attentive reader (neuronal)  
[Chen et al., 2016]

**GA** gated-attention reader (neuronal)  
[Dhingra et al., 2017]

**CPerf** % de questions avec réponses atteignables (échantillon de 500 questions)
WDW [Onishi et al., 2016]

- 200k questions à trou (cloze) à choix multiples construites sur le [Gigaword newswire](https://tticnlp.github.io/who_did_what/) corpus

- processus pour sélectionner des paires \((q, p, c)\)
  1. choisir \(q\) la première phrase d'une nouvelle
  2. considérer ses entités nommées de \(q\), une sera retirée
  3. identifier une autre nouvelle \(p\) ”proche” via Lucene (question : \(q\))
  4. identifier les entités nommées présentes dans \(p\) comme réponse potentielle \((c)\)

- plus filtrage, retrait de questions trop faciles

- corpus, leaderbord,
  https://tticnlp.github.io/who_did_what/
passage: Britain’s decision on Thursday to drop extradition proceedings against Gen. Augusto Pinochet and allow him to return to Chile is understandably frustrating ... Jack Straw, the home secretary, said the 84-year-old former dictator’s ability to understand the charges against him and to direct his defense had been seriously impaired by a series of strokes. ... Chile’s president-elect, Ricardo Lagos, has wisely pledged to let justice run its course. But the outgoing government of President Eduardo Frei is pushing a constitutional reform that would allow Pinochet to step down from the Senate and retain parliamentary immunity from prosecution. ...

question: Sources close to the presidential palace said that Fujimori declined at the last moment to leave the country and instead he will send a high level delegation to the ceremony, at which Chilean President Eduardo Frei will pass the mandate to XXX.

choix: (1) Augusto Pinochet (2) Jack Straw (3) Ricardo Lagos
## WDW  [Onishi et al., 2016]

<table>
<thead>
<tr>
<th>System</th>
<th>WDW</th>
<th>CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word overlap</td>
<td>0.47</td>
<td>–</td>
</tr>
<tr>
<td>Sliding window</td>
<td>0.48</td>
<td>–</td>
</tr>
<tr>
<td>Distance</td>
<td>0.46</td>
<td>–</td>
</tr>
<tr>
<td>Sliding window + Distance</td>
<td>0.51</td>
<td>–</td>
</tr>
<tr>
<td>Semantic features</td>
<td>0.52</td>
<td>–</td>
</tr>
<tr>
<td>Attentive Reader</td>
<td>0.53</td>
<td>0.63&lt;sup&gt;I&lt;/sup&gt;</td>
</tr>
<tr>
<td>Attentive Reader (relaxed train)</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>Stanford Reader</td>
<td>0.64</td>
<td>0.73&lt;sup&gt;II&lt;/sup&gt;</td>
</tr>
<tr>
<td>Stanford Reader (relaxed train)</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>AS Reader</td>
<td>0.57</td>
<td>0.70&lt;sup&gt;III&lt;/sup&gt;</td>
</tr>
<tr>
<td>AS Reader (relaxed train)</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>GA Reader</td>
<td>0.57</td>
<td>0.74&lt;sup&gt;IV&lt;/sup&gt;</td>
</tr>
<tr>
<td>GA Reader (relaxed train)</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td><strong>Human Performance</strong></td>
<td><strong>84/100</strong></td>
<td><strong>0.75+&lt;sup&gt;II&lt;/sup&gt;</strong></td>
</tr>
</tbody>
</table>

**Table 4:** System performance on test set. Human performance was computed by two annotators on a sample of 100 questions. Result marked *I* is from (Hermann et al., 2015), results marked *II* are from Chen et al., 2016.
A study on similarity and relatedness using distributional and wordnet-based approaches.

A thorough examination of the cnn/daily mail reading comprehension task.
In 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), pages 2358–2367.

Think you have solved question answering? try arc, the AI2 reasoning challenge.
CoRR.

