Université de Montréal at the NTCIR-11 IMine task

Arbi BOUCHOUCHA, Jian-Yun NIE and Xiaohua LIU

Dept. of Computer Science and Operations Research

- December 2014 -

Our Participation

- We participate in two sub-tasks:
  - **Subtopic Mining** (English)
  - **Document Ranking** (English)

Proposed Approach (1/2)

- Embedding framework for learning latent aspect vectors of the query

- Each aspect vector is mapped by an expansion term

- Candidate expansion terms are selected from different resources (E.g. ConceptNet, Wikipedia, query logs, etc)

- **Our Goal**: Make two known similar terms similar whatever the resource used to recognize the similarity between them

  - E.g. 'program' and 'algorithm' for the query "java"

Proposed Approach (2/2)

- Objective function:

$$\min \sum_{r \in R} \sum_{e_i, e_j \in E, i \neq j} \frac{1}{2} \, . \, w_r \, . \left(sim\left(\overrightarrow{e_i}, \overrightarrow{e_j}\right) - sim_r\left(e_i, e_j\right) \right)^2$$

subject to $\quad ||\vec{e}||_2^2 \leq 1, e^k \geq 0, k = 1,2, \dots, N, \forall e \in E$

- Maximal Marginal Relevance-based Expansion (MMRE):

(Bouchoucha et al. 2014) ‡

$$\overrightarrow{e^*} = argmax_{e \in E - ES} \left(\beta \, . sim(\vec{e}, \vec{q}) - (1 - \beta) \, . \, max_{\overrightarrow{e'} \in ES} \, sim\left(\vec{e}, \overrightarrow{e'}\right) \right)$$

where $\quad sim\left(\overrightarrow{e_i}, \overrightarrow{e_j}\right) = \overrightarrow{e_i} . \overrightarrow{e_j} = \sum_{k=1,\dots,N} e_i^k \, . \, e_j^k$

‡ A. Bouchoucha, X. Liu, and J.-Y. Nie. "Integrating Multiple Resources for Diversified Query Expansion". In *Proceedings of ECIR*, pp. 437-442, Amsterdam, Netherlands, 2014.

Resource-based Similarity Functions

- Wikipedia, ConceptNet, query logs and documents feedback: same definitions as in (Bouchoucha et al. 2014) ‡

- Query suggestions: $$sim_{QS}(e_i, e_j) = \frac{2 \,.\, n(e_i, e_j)}{n(e_i) + n(e_j)}$$

$n(e_i)$ (resp. $n(e_j)$) : number of times term e_i (resp. e_j) appears in query suggestions, and $n(e_i, e_j)$ is the number of times when both expansion terms appear in the suggestions of the same query.

‡ A. Bouchoucha, X. Liu, and J.-Y. Nie. "Integrating Multiple Resources for Diversified Query Expansion". In *Proceedings of ECIR*, pp. 437-442, Amsterdam, Netherlands, 2014.

Université de Montréal

Subtask 1: Subtopic Mining (1/2)

- **Query Classification:** 33 features derived from the resources

- 3 target classes:

    - **Ambiguous**
    - **Broad**
    - **Clear**

Subtask 1: Subtopic Mining (1/2)

Feature	Description	Total
***** Query-dependent:***		
`NumTerms`	Number of terms of q after removing stopwords	1
`Quest`	Whether q is a question?	1
***** Query-independent:***		
** Feedback documents-based:*		
`SongEtAl`	The set of 11 features described in [19] (excepting the feature about the number of terms in q)	11
`AvgPMI`	Average mutual information score between the terms of q and the top 10 terms that co-occur a lot with the terms of q in D	1
`ClarityScore`	Clarity score of q computed on D and the whole collection [9]	1
** Wikipedia-based:*		
`NumInterp`	Number of (possible) interpretations of q in the Wikipedia disambiguation page of q	1
`WikiLength`	Wikipedia page length (number of *different* words) that matches with q	1
** Query logs-based:*		
`NumClicks`	Max, Min and average number of clicked URLs for q in all the sessions	3
`PercentageClicks`	Percentage of shared clicked URLs between different users who issued q	1
`ClickEntropy`	Click entropy of the query q [11]	1
`NumSessions`	Total number of sessions with q	1
`SessionLength`	Max, Min and average session duration (in seconds) with q	3
`NumTermsReform`	Total number of different terms added by users to reformulate q in all the sessions	1
`ReformLength`	Max, Min and average number of terms added by users to reformulate q in all the sessions	3
** ConceptNet-based:*		
`NumDiffNodes`	Number of different adjacent nodes that are related to the nodes of the graph of q	1
`AvgCommonNodes`	Average number of common nodes shared between the nodes of the graph of q (*i.e.* nodes that are connected to at least two edges)	1
`NumDiffRelations`	Number of different relation types defined between the adjacent nodes in the graph of q	1
Grand Total		**33**

Université de Montréal

Subtask 1: Subtopic Mining (2/2)

- **Query Disambiguation and Predicting Subtopic Importance:**

  • Disambiguate each ambiguous query ➔ generate the first-level subtopics (sub-queries or query interpretations)

  • Use Wikipedia disambiguation page + query logs (query reformulations, user sessions)

  • Importance of i^th sub-query: $score_q(i) = \frac{n_q(i)}{|E_q|}$

  where E_q is the set of terms added to reformulate *q*, and $n_q(i)$ is the number of terms from E_q that appear in the Wikipedia disambiguation page of *q*.

Subtask 2: Document Ranking

- **Selective diversification:** Different queries (ambiguous/broad/clear) require different diversification strategies

- **Clear queries**: No need for diversification (results of the baseline)
- **Broad queries**: Search results of the expanded query (expansion terms are those obtained by the embedding framework)
- **Ambiguous queries**: Greed selection of documents from different sets

$$d^* = \arg\max_{d \in \cup D_i - S} \left(\frac{rel(d).score_q(i)}{rank(d)} \right)$$

where D_i is the search results of the i[th] sub-query of q and S is the set of documents already selected.

Submitted Runs

- Three runs: (for each subtask)
  - Embedding Framework with an explicit modeling of query aspects

- **1st run**: (UM13-S-E-1A ; UM13-D-E-1A) – five resources: Wikipedia, ConceptNet, query logs, feedback documents , and query suggestions.

- **2nd run**: (UM13-S-E-2A ; UM13-D-E-2A) – four resources: Wikipedia, ConceptNet, query logs, and feedback documents.

- **3rd run**: (UM13-S-E-3A ; UM13-D-E-3A) – one resource: Query logs.

Université de Montréal

Experimental Setting

- **Used resources:**
  - English Wikipedia dumps of July 8th, 2013
  - MSN query logs of 2006
  - ConceptNet 5
  - Top 50 feedback documents
  - Query suggestions from Bing, Google and Yahoo!

- **Query classification:**
  - Use SVM-Light tool for non-linear SVM (with RBF kernel):
    http://svmlight.joachims.org
  - 450 training queries publicly available in
    http://www.ccc.ipt.pt/~ricardo/datasets/GISQC_DS.html

Results: Query Classification

Query class	*Precision*	*Recall*	*F1-measure*
Ambiguous (a)	75.00%	56.30%	64.30%
Broad (b)	44.83%	76.47%	56.52%
Clear (c)	88.89%	47.06%	61.54%

- Average classification success: 60% of the queries
- Failure to distinguish between broad and clear queries
  - Possible reason: Aspects of a broad query tend to overlap and most of the intents tend to be almost similar

➔ Similar behaviour as clear queries

Results: Subtopic Mining

runs	Hscore	Fscore	Sscore	H-measure
1st run	0.2056	**0.1624**	**0.0059**	0.0047
2nd run	**0.2064**	0.1624	0.0059	**0.0049**
3rd run	0.1766	**0.1624**	0.0049	0.0037

- 1st and 2nd run lead to very comparable results

➔ Query suggestions seems to be not helpful to improve search results! (need further investigation in the future to confirm …)

- 1st and 2nd run vs. 3rd run

➔ Using multiple resources yields to improved results than using a single one.

Results: Subtopic Mining


```
<topic number="61" type="ambiguous">
  windows software;1;0.64;0;update;1;0.96;
  windows software;1;0.64;0;installer;2;0.91;
  windows software;1;0.64;0;8;3;0.89;
  windows software;1;0.64;0;versions;4;0.82;
  windows software;1;0.64;0;license;5;0.75;
  windows software;1;0.64;0;defender;6;0.63;
  windows software;1;0.64;0;replacement;7;0.61;
  windows software;1;0.64;0;recovery;8;0.54;
  windows software;1;0.64;0;vista;9;0.41;
  windows software;1;0.64;0;live;10;0.29;
  windows house;2;0.35;0;catalog;1;0.90;
  windows house;2;0.35;0;treatment;2;0.82;
  windows house;2;0.35;0;glass;3;0.81;
  windows house;2;0.35;0;paint;4;0.78;
  windows house;2;0.35;0;pictures;5;0.77;
  windows house;2;0.35;0;construction;6;0.63;
  windows house;2;0.35;0;sizes;7;0.55;
  windows house;2;0.35;0;tinting;8;0.49;
  windows house;2;0.35;0;manufacturer;9;0.31;
  windows house;2;0.35;0;pulls;10;0.27;
</topic>
```

Results: Document Ranking

runs	Coarse-grain	Fine-grain
1st run	**0.6254**	**0.5566**
2nd run	0.6001	0.5309
3rd run	0.4474	0.3770

Coarse-grain (1st level subtopic) results

runs	AP	RBP	nDCG	ERR	I-rec	D#-nDCG
1st run	**0.5479**	**0.1655**	**0.5108**	0.4236	**0.7899**	**0.6511**
2nd run	0.4782	0.1489	0.4750	**0.4251**	0.7520	0.6137
3rd run	0.2520	0.1025	0.3162	0.2880	0.5692	0.4397

Fine-grain (2nd level subtopic) results

runs	AP	RBP	nDCG	ERR	I-rec	D#-nDCG
1st run	**0.5479**	**0.1480**	**0.4629**	**0.2628**	**0.6310**	**0.5469**
2nd run	0.4782	0.1340	0.4301	0.2602	0.5874	0.5089
3rd run	0.2520	0.0915	0.2901	0.1807	0.3798	0.3331

- Combining multiple resources yields to a better coverage of query aspects.

Conclusions

- We experimented a new approach for diversified query expansion based on embedding

- Our approach learns query aspects by selecting (good) expansion terms that cover the query aspects

- Several different resources are used

- Promising results!

- Our best document ranking run is ranked **No. 2 of all 15 runs** in terms of coarse-grain and fine-grain results

Thank you.
Questions?