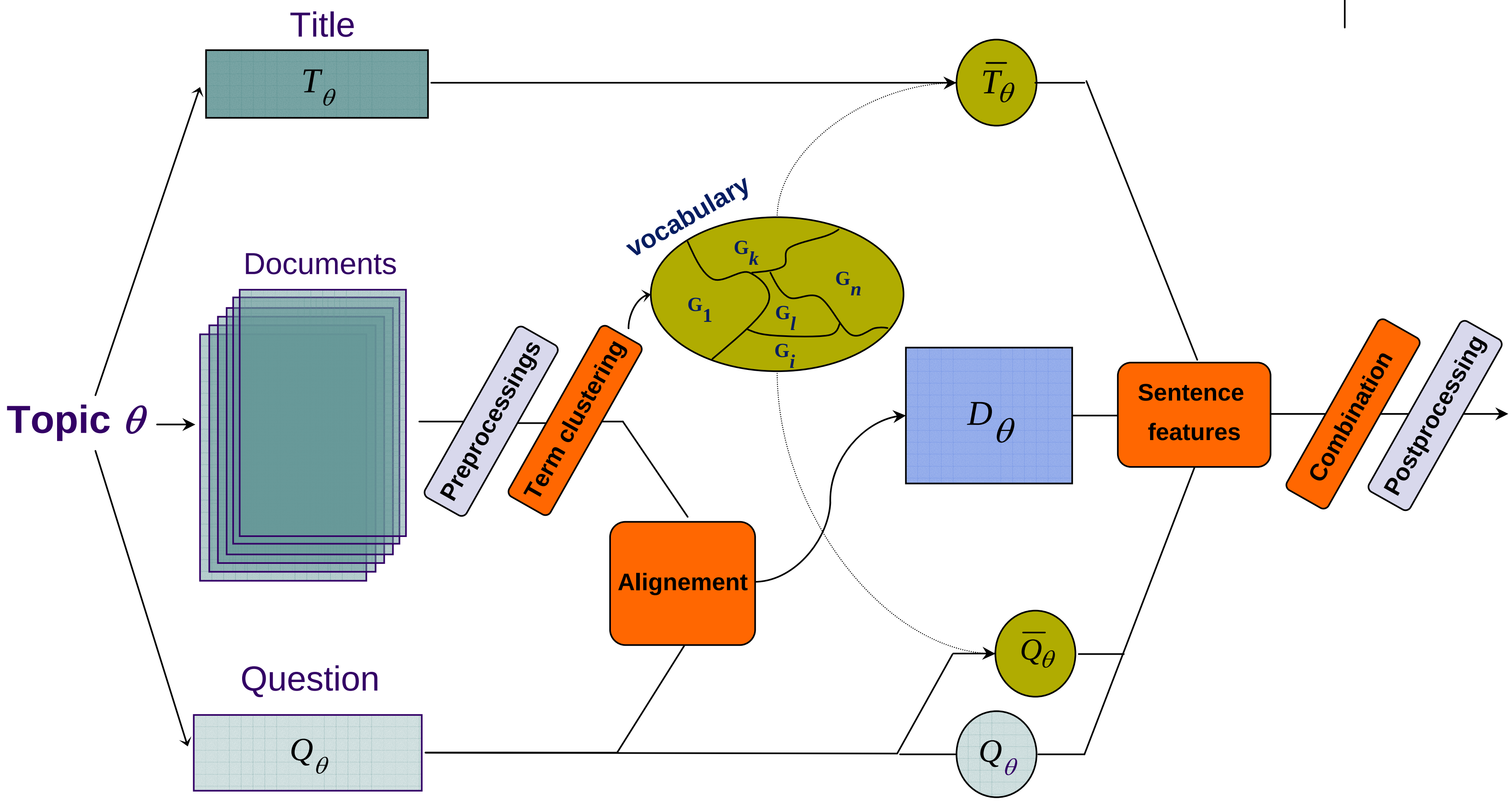
A Contextual Query Expansion Approach by Term Clustering for Robust Text Summarization

Massih Amini and Nicolas Usunier

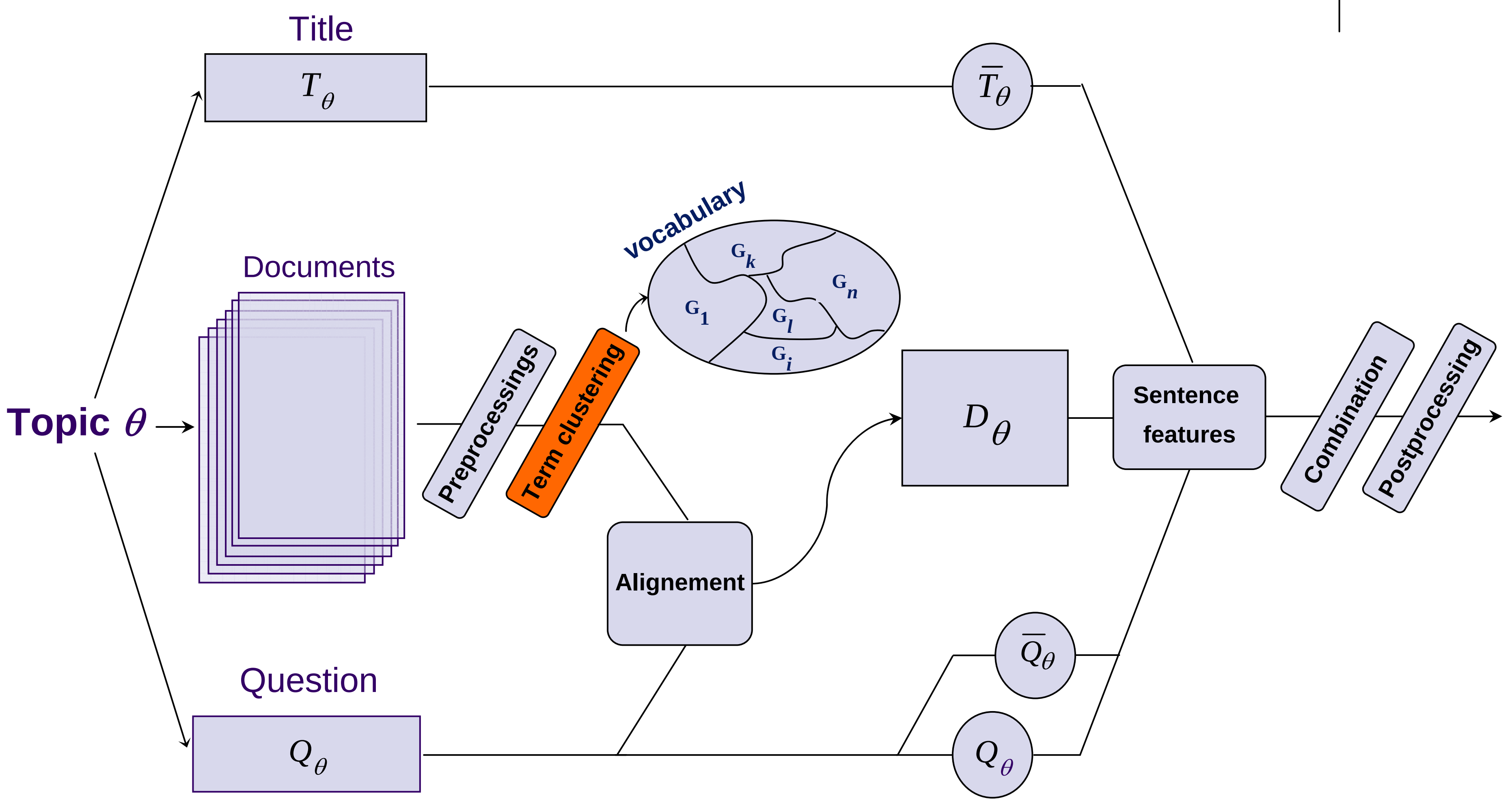
April the 26th 2007

Université Pierre et Marie Curie (Paris 6)
Laboratoire d'Informatique de Paris 6

LIP6 summarizer



Term clustering



Term clustering

- Hypotheses:
  - Words occurring in the same context with the same frequency are topically related (*context* $\equiv$ *document*),
  - Each term is generated by a mixture density,

$$p(\vec{w}|\Theta) = \sum_{k=1}^{K} \pi_k \, p(\vec{w}|c = k, \theta_k)$$

  - Each term of the vocabulary V belongs to one and only one term cluster $\rightarrow$ to each term w_i we associate an indicator vector class $t_i = \{t_{hi}\}_h$

$$\forall w_i \in V, \forall k, y_i = k \Leftrightarrow t_{ki} = 1 \text{ and } \forall h \neq k, t_{hi} = 0$$

Term clustering (2)

- Each vocabulary term *w* is represented as a bag-of-documents:

$$\vec{w} = \left\langle tf\left(w, d_i\right)\right\rangle_{i \in \{1,\ldots,n\}}$$

- Term clustering is performed using the CEM algorithm.

Term clustering (3): CEM algorithm

- **Input:**
  - An initial partition $C^{(0)}$ is chosen at random and the class conditional probabilities are estimated on the corresponding classes
- **Repeat until convergence of the complete data log-likelihood:**
  - **E-step:** Estimate the posterior class probability that each term w_j belongs to C_k,
  - **C-step:** Assign each term probability with maximal posterior probability according to the previous step,
  - **M-step:** Estimate the new mixture parameters which maximize the complete data log-likelihood
- **Output:** Term clusters.

Term clustering (4): examples

D0714: Term cluster containing *Napster*

digital trade act format drives allowed illegally net **napster** search stored alleged released musical electronic internet signed intended idea billions distribution exchange mp3 music songs tool

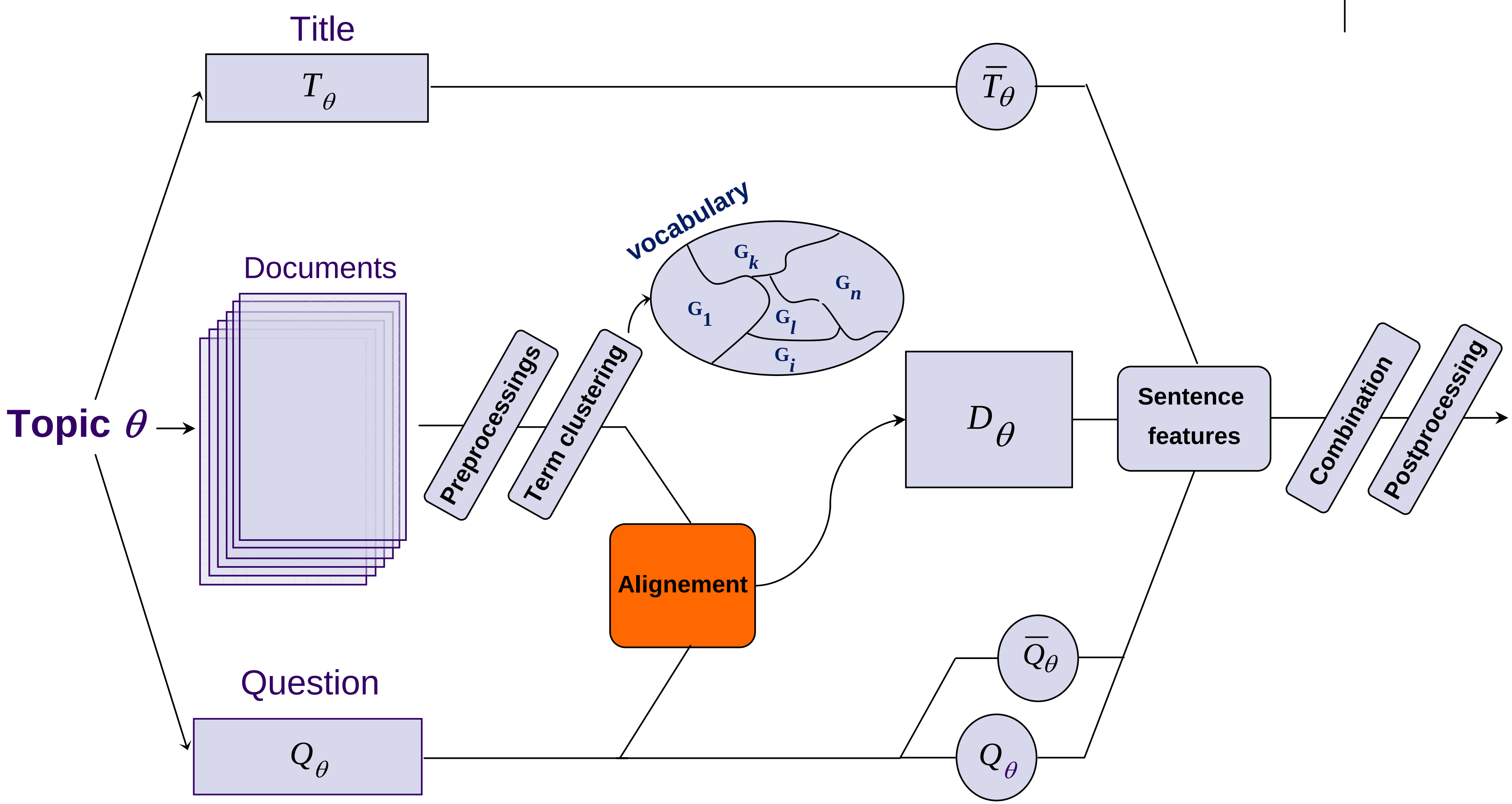
D0728: Term cluster containing *Interferon*

depression **interferon** antiviral protein drug ribavirin combination people hepatitis liver disease treatment called doctors cancer epidemic flu fever schering plough corp

D0705: Term cluster containing *basque* and *separatism*

basque people separatist armed region spain **separatism** eta independence police france batasuna nationalists herri bilbao killed

Sentence alignment



Sentence alignment

- **Aim**: Remove non-informative sentences of each topic (those which do not likely contain the answer to the topic question).

- **Hypothesis**: Sentences containing the answer to the topic question are those which have the maximal semantic similarity with the question.

- **Tool**: Marcu's alignment algorithm (Marcu 99).

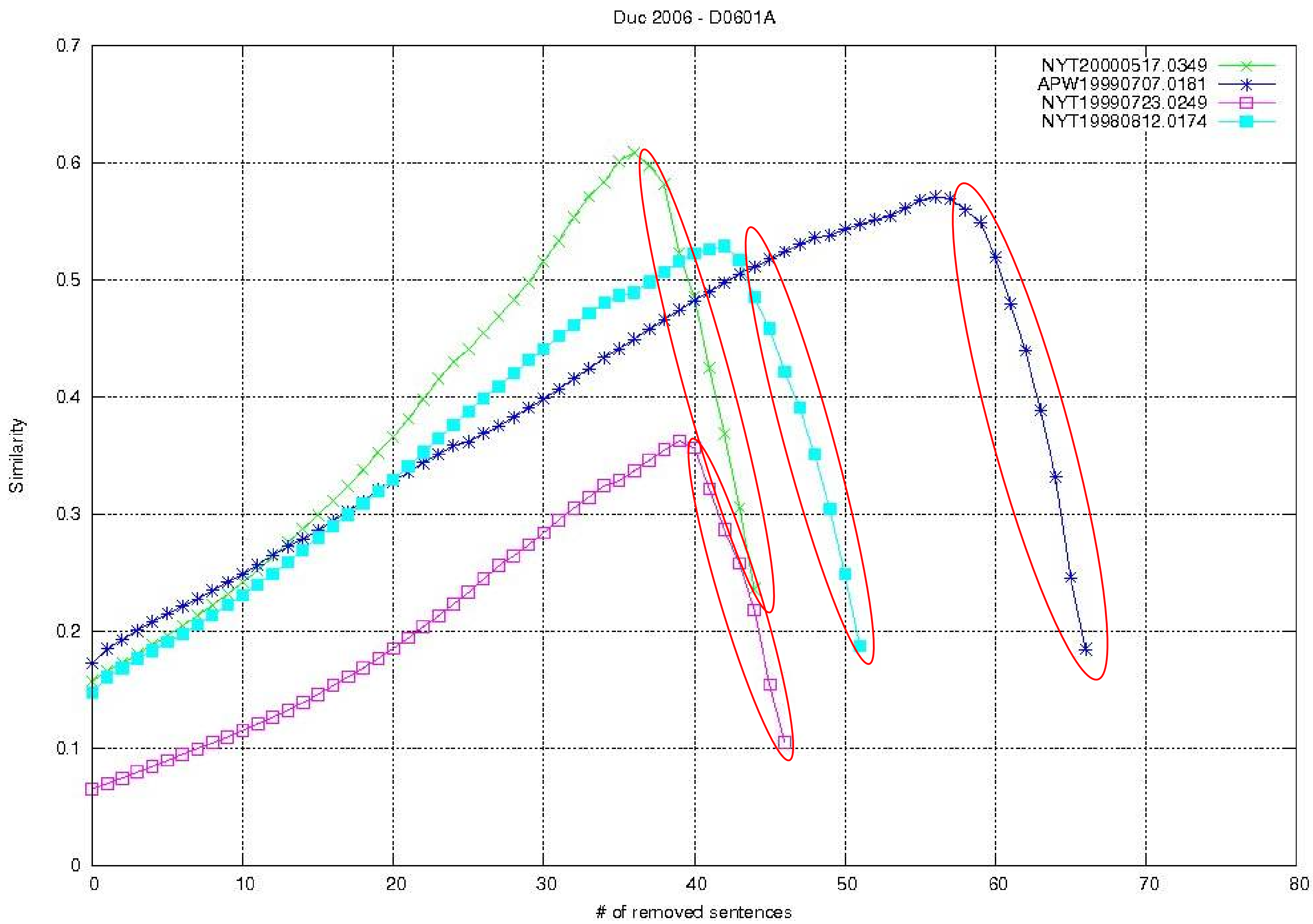
Sentence alignment: the algorithm (2)

- **Input: topic question and a document**
- **Repeat until the similarity of the remaining document set decreases**
  - **Remove the sentence from the current set such that its removal maximizes the similarity between the question and the rest of the sentences**
- **Output: The set of candidate sentences**

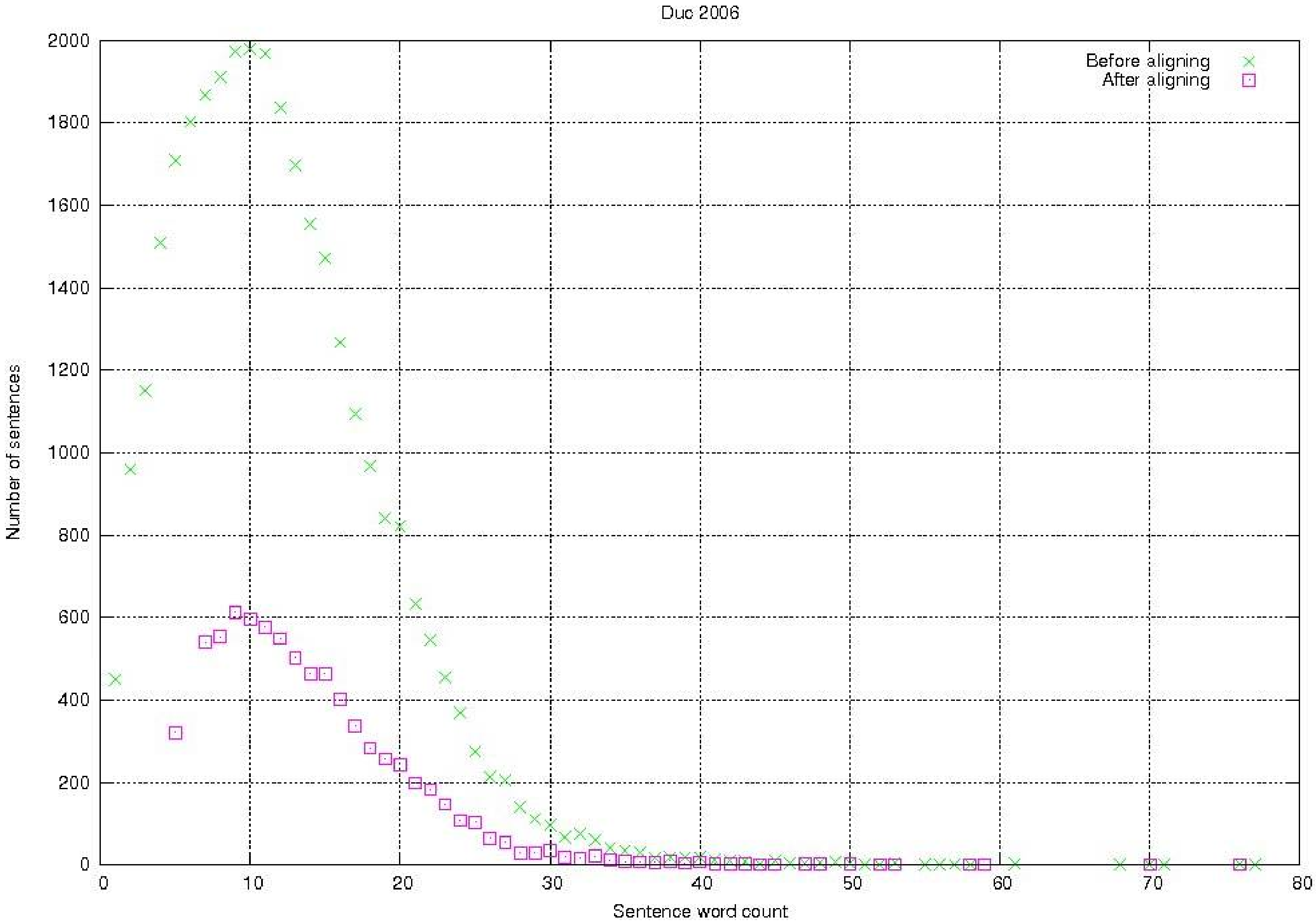
$$Sim(S,Q) = \frac{\sum_{w \in S \cap Q} c(w,S)\, c(w,Q)}{\sum_{w \in S} c^2(w,S) \sum_{w \in Q} c^2(w,Q)}$$

$$c(w,Z) = tf(w,Z) \times log(df(w))$$

Sentence alignment: the behavior (3)



Sentence alignment: filtered word distribution (4)



Remaining sentences in some documents of topic D0708

Question$_{D0708}$: What countries are having chronic potable water shortages and why?

Document: XIE19970212.0042

Before

The Addis Ababa Regional Water and Sewerage Authority announced that the shortage of potable water in the capital city of Ethiopia will be solved in the last quarter of this year.

According to a report here today, the announcement was made by Tadesse Kebede, general manager of the authority.

Currently, the authority supplies only 60 percent of the city's potable water demand.

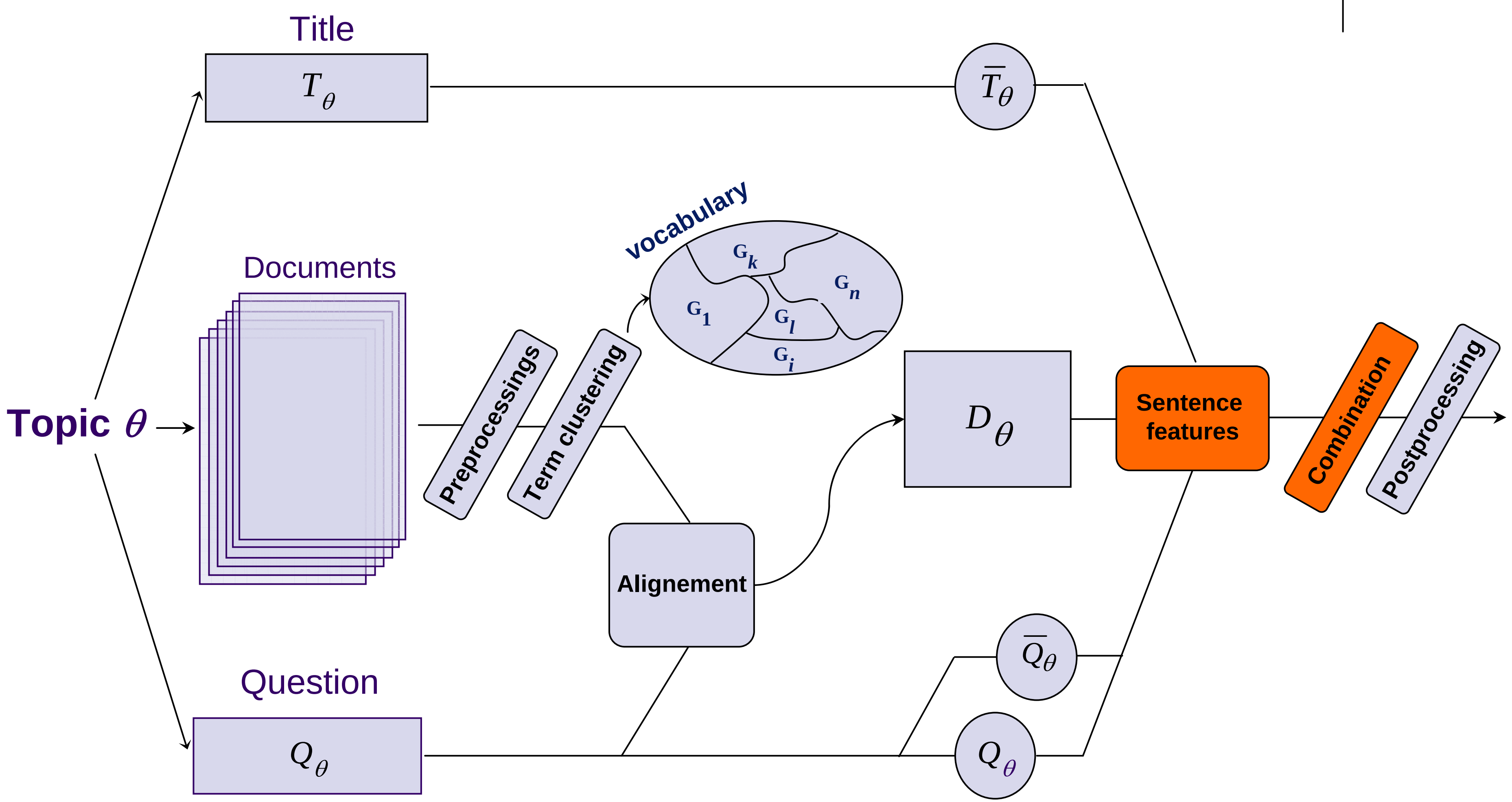
Tadesse said 18 water supply projects are underway at various stages, adding that one of such projects involved the sinking of 25 wells at Akaki, about 20 kilometers from Addis Ababa, which will supply 75,000 cubic meters of water daily to the capital city.

After

The Addis Ababa Regional Water and Sewerage Authority announced that the shortage of potable water in the capital city of Ethiopia will be solved in the last quarter of this year.

Tadesse said 18 water supply projects are underway at various stages, adding that one of such projects involved the sinking of 25 wells at Akaki, about 20 kilometers from Addis Ababa, which will supply 75,000 cubic meters of water daily to the capital city.

Sentence features and combination



Sentence features

- From the topic title T_θ and question Q_θ we derived 3 queries:
  - q_1 = question keywords,
  - q_2 = question keywords expanded with their word clusters,
  - q_3 = title keywords expanded with their word clusters,
- Features

Feature	Query	Score
F_1	q_1	$common\_terms(q_1, s)$
F_2	q_1	$cosine(q_1, s)$
F_3	q_2	$ldf(q_2, s)$
F_4	q_3	$ldf(q_3, s)$

Combination: why?

- Spearman rank order correlation

Object	**Rank Sys1**	**Rank Sys2**
1	r_1	s_1
2	r_2	s_2
.	.	.
.	.	.
.	.	.
n	r_n	s_n

$$CorrSpearman(Sys_1, Sys_2) = \frac{Cov(r,s)}{\sigma_r \sigma_s} = 1 - \frac{6\sum_{i=1}^{n}(r_i - s_i)^2}{n(n^2-1)}$$

Features	F_1	F_2	F_3	F_4
F_1	$*$	0.198	0.186	0.141
F_2	$*$	$*$	0.095	0.086
F_3	$*$	$*$	$*$	0.123

Combination by learning

- We have developed a learning based ranking model for extractive summarization.

  ☞ Amini M.-R., Tombros A., Usunier N., Lalmas M. *Learning Based Summarization of XML Documents.* Journal of Information Retrieval (2007), to appear.

- For learning we need a training set where for each sentence of each topic a label class is available,

- We constructed a training set by labeling sentences having highest Rouge2 Average-F measure as relevant sentences to the summary.

This strategy sounds good but it doesn't work.

Handcrafted weighted

- We also tried to fusion ranked lists obtained from each feature using the weighted borda fuse algorithm (Aslam et Montague, 2001).

This strategy didn't work either.

- We determined combination weights for which we obtained the best Rouge2 Average F-measure on Duc2006.

Results

Average F of Rouge-2

DUC 2007			
Id	Mean	95% low. C.I.	95% upp. C.I.
D	0.17175	0.15322	0.19127
C	0.14993	0.13372	0.16741
J	0.14141	0.12265	0.16274
G	0.13903	0.12312	0.15385
E	0.13764	0.12413	0.15315
B	0.13740	0.11372	0.16061
F	0.13739	0.12097	0.15530
A	0.13430	0.11765	0.15108
I	0.13328	0.11017	0.15481
H	0.12702	0.11448	0.13995
15	0.12285	0.11800	0.12768
4	**0.11886**	**0.11467**	**0.12351**
29	0.11725	0.11245	0.12225
24	0.11605	0.11040	0.12133

Results (2)

Average F of Rouge-SU4

DUC 2007			
System Id	Mean	95% lower condifence intervals	95% upper condifence intervals
D	0.21461	0.20154	0.22922
C	0.19846	0.18350	0.21478
J	0.19378	0.17834	0.21139
E	0.19266	0.18147	0.20490
F	0.19165	0.17905	0.20506
A	0.18902	0.17749	0.20182
G	0.18761	0.17638	0.19886
B	0.18620	0.16685	0.20543
H	0.18044	0.17067	0.18967
I	0.18016	0.16292	0.19648
15	0.17470	0.16997	0.17939
24	0.17304	0.16800	0.17769
4	**0.17007**	**0.16646**	**0.17381**
29	0.16635	0.16163	0.17113

Conclusion

- Query expansion by term clustering may help to simply resolve complex NLP problems,
- Combination of features showed promising results,
- It would be worth to constitute training sets (for example making models by extracting manually sentences for summaries)

Thank you