Geo-temporal Information Retrieval Based on Semantic Role Labeling and Rank Aggregation

Yoonjae Jeong, Gwan Jang, Kyung-min Kim, and Sung-Hyon Myaeng
Korea Advanced Institute of Science and Technology
291 Daehak-ro (373-1 Guseong-dong), Yuseong-gu, Daejeon 305-701, Republic of Korea
{hybris, gjang, kimdarwin, myaeng}@kaist.ac.kr

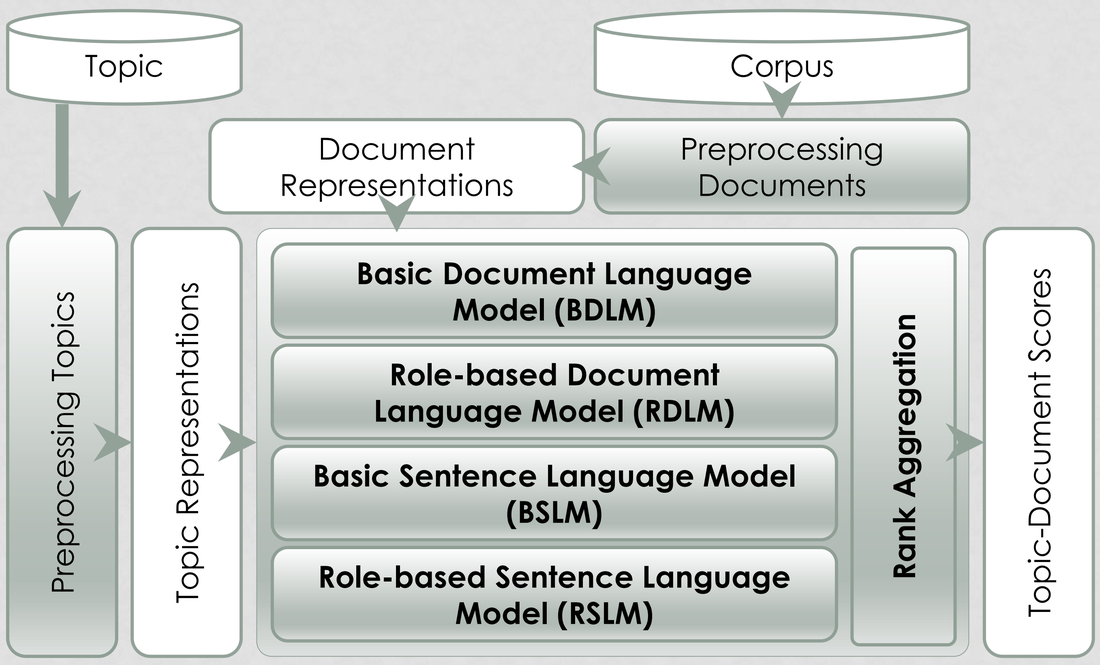
1. Introduction

NTCIR-9 GeoTime Task is about geographic and temporal search in news articles. We participated in the English sub-task only. A topic requests the information on where and when a particular event occurred or what event happened at a specific time and a location.

Our basic idea is to add locational and temporal aspects to terms in a document using Semantic Role Labeling (SRL). A semantic role is the underlying relationship that a participant (linguistic constituent) has with the main verb in a clause.

2. The Proposed Method

- Both documents and topics are processed for SRL and represented with terms and their semantic roles. Topics are processed further to identify the question types.
- For the purpose of matching topic and document representations, we propose four variations of language modeling.



Overview of the proposed geo-temporal information retrieval

- The four retrieval models (BDLM, RDLM, BSLM, and RSLM) have different advantages and disadvantages.
- Based on the observation, a method for combining the ranked lists of retrieved documents from the four models is devised.
- The rank aggregation module in the system assigns the final scores to the retrieved documents.

(1) Document Representations

Attribute	Description
T_V	A set of verb in document
T_A	A set of terms with numbered argument roles (A0-5) in document
$T_{AM\text{-}LOC}$	A set of terms with location (AM-LOC) roles in document
$T_{AM\text{-}TMP}$	A set of terms with temporal role (AM-TMP) in document

(2) Topic Representations

Attribute		Description
Question types	Q-LOC	Whether a question is about location or not.
	Q-TMP	Whether a question is about time or not.
	Q-AGT	Whether is a question about agent or not.
	Q-MSC	The others
Set of terms	V_V	A set of vocabularies in verb role in topic
	V_A	A set of vocabularies in numbered argument (A0-5) roles in topic
	$V_{AM\text{-}LOC}$	A set of vocabularies in locational role (AM-LOC) in topic
	$V_{AM\text{-}TMP}$	A set of vocabularies in temporal role (AM-TMP) in topic

(3) Information Retrieval Models

- **Basic Document Language Model (BDLM).** A basic language model using the Dirichlet smoothing method
- **Role-based Document Language Model (RDLM).** Base on the document representation, we built the language models for their attributes. The RDLM combines those language models.
- **Basic Sentence Language Model (BSLM).** Sometimes the relevant information related to a topic is fully contained in one sentence in document. BSLM calculate the scores for sentences.
- **Role-based Sentence Language Model (RSLM).** RSLM adds semantic roles to BSLM in the same way RDLM was constructed out of BDLM.

(4) Rank Aggregation

RSLM reveals the property of information extraction while BDLM shows the nature of general information retrieval.

Relevant Documents	Rank for GeoTime-0025			
	BDLM	RDLM	BSLM	RSLM
NYT_ENG_20041226.0096	24	201	251	**4**
NYT_ENG_20041229.0208	167	322	133	**3**
NYT_ENG_20041230.0186	**3**	18	298	102
NYT_ENG_20041230.0204	**8**	26	302	98
NYT_ENG_20041230.0245	**4**	17	299	101
NYT_ENG_20041230.0256	**6**	25	303	100
NYT_ENG_20041231.0009	**2**	19	300	99
NYT_ENG_20050328.0205	36	349	88	162

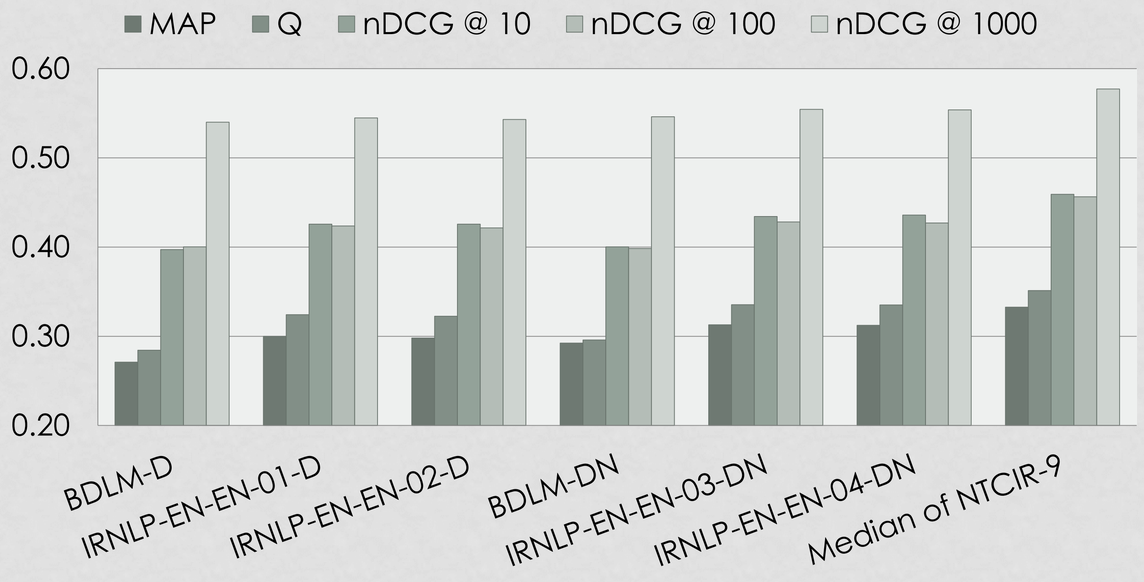
- **The rank aggregation based on Markov Chain (Dwork, et al., 2001)**
  - We adopted Dwork, et al.'s MC2 heuristic because it is arguably the most representative of minority viewpoints of sufficient statistical significance; it protects specialist views.
- **The proposed rank aggregation with threshold**
  - The effective ranks for aggregations of RDLM, BSLM, and RSLM are a small number of top ones.
  - We applied the threshold θ to the elements of transition matrix for those models.

3. Evaluation

(1) Description of Runs in NTCIR-9

RUN	Topic Source	Aggregation	Aggregation Threshold (θ)
IRNLP-EN-EN-1-D	Description only	(BD, RD, BS, & RS) LM	150
IRNLP-EN-EN-2-D	Description only	(BD, RD, BS, & RS) LM	200
IRNLP-EN-EN-3-DN	Description & Narrative	(BD, RD, BS, & RS) LM	200
IRNLP-EN-EN-4-DN	Description & Narrative	(BD, RD, BS, & RS) LM	150
BDLM-D	Description only	BDLM	-
BDLM-DN	Description & Narrative	BDLM	-

(2) Results from NTCIR-9



- **Topics showing high performances have:**
  - Verbs related to the activities or states of agents clearly (e.g. "murder", "hijack", "kill", and so on).
  - The terms that are also not ambiguous: proper nouns or very specific number of theme (e.g. "4 people" in GeoTime-0033).
- **Topics showing low performances have:**
  - Many errors in the analysis of topics
  - The verbs related to the existence or occurrence of agent or theme. (e.g., "occur", "happen")
  - They sometimes require inference or term expansion.

4. Conclusion

- A new geo-temporal information retrieval method that utilizes semantic role labeling (SRL) and rank aggregation.
- While the SRL-based method is not always superior, they complement the usual language modeling and warrant the proposed rank aggregation method.
- Through an analysis of the result, we found that the term expansion and weighting are necessary.