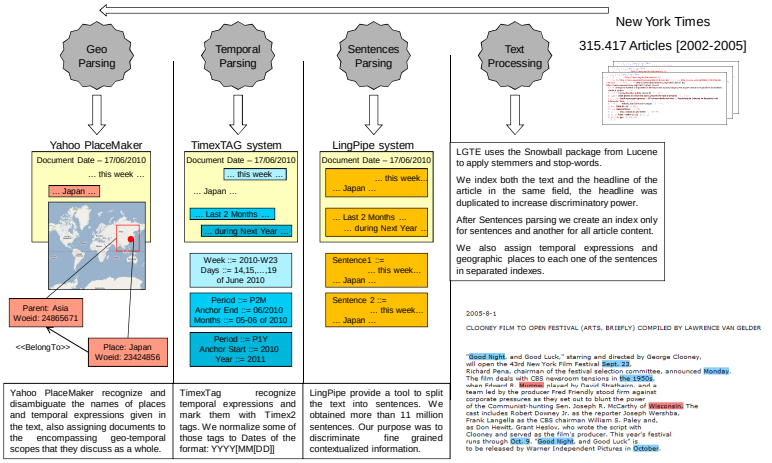
Experiments with Geo-Temporal Expressions filtering and query expansion at document and phrase context

Jorge Machado, José Borbinha, Bruno Martins

GTE @NTCIR

We describe an evaluation experiment on GeoTemporal Document Retrieval created for the GeoTime evaluation task of NTCIR 2010. GeoTemporal Retrieval aims at to improve retrieval results using Geographic and Temporal dimensions of relevance. To accomplish that task, systems need to extract geographic and temporal information from the documents, and then explore semantic relations among those dimensions within the documents. Since this is the first time the task is taking place our aim is to evaluate some basic techniques in order to set some research directions of our work. We aim to understand the relevance of temporal and geographic expressions for filtering purposes. The geographic expressions were extracted with Yahoo PlaceMaker and for temporal expressions we used the TIMEXTAG system. We experimented techniques using both the overall document and sentence resolutions, as also one mixed approach. We also used a query expansion mechanism in topics with no filters defined. We used the BM25 as retrieval model and preprocessed the topics with a semi-automatic methodology to create structures that let us create our filters and expansions. We learned that the sentence level is not a very good approach (but we got clues that probably the paragraph context resolution could improve the results) and the geographic and temporal expressions base filters had shown good performance.

Collection Processing

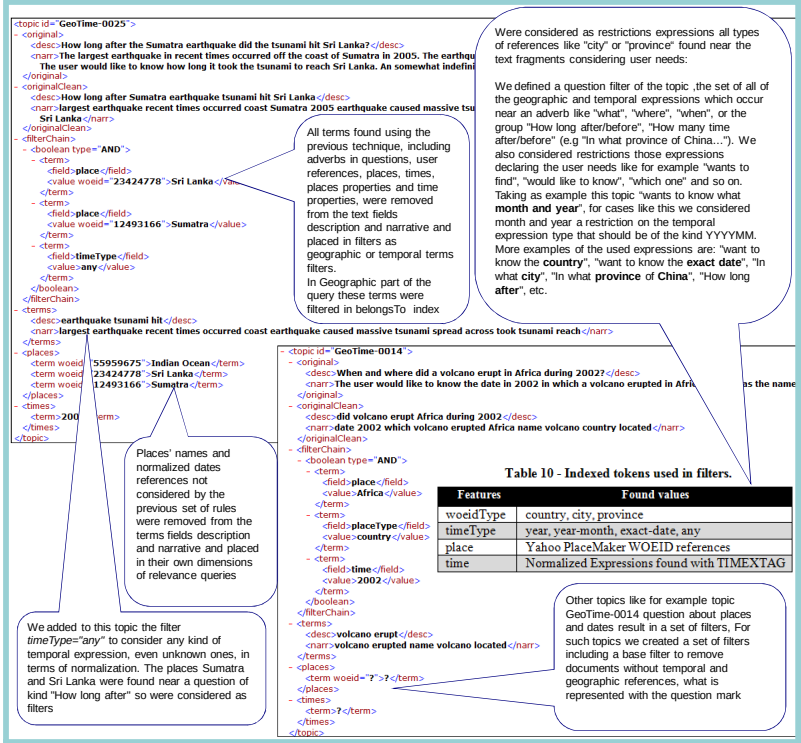


Yahoo PlaceMaker recognize and disambiguate the names of places and temporal expressions given in the text, also assigning documents to the encompassing geo-temporal scopes that they discuss as a whole.

TimexTag recognize temporal expressions and mark them with Timex2 tags. We normalize some of those tags to Dates of the format: YYYY[MM[DD]]

LingPipe provide a tool to split the text into sentences. We obtained more than 11 million sentences. Our purpose was to discriminate fine grained contextualized information.

Topics Processing



All terms found using the previous technique, including adverbs in questions, user references, places, times, places properties and time properties, were removed from the text fields description and narrative and placed in filters as geographic or temporal terms filters.
In Geographic part of the query these terms were filtered in belongsTo index

Were considered as restrictions expressions all types of references like "city" or "province" found near the text fragments considering user needs:

We defined a question filter of the topic ,the set of all of the geographic and temporal expressions which occur near an adverb like "what", "where", "when", or the group "How long after/before", "How many time after/before" (e.g "In what province of China…"). We also considered restrictions those expressions declaring the user needs like for example "wants to find", "would like to know", "which one" and so on. Taking as example this topic "wants to know what **month and year**", for cases like this we considered month and year a restriction on the temporal expression type that should be of the kind YYYYMM. More examples of the used expressions are: "want to know the **country**", "want to know the **exact date**", "In what **city**", "In what **province** of **China**", "How long **after**", etc.

Places' names and normalized dates references not considered by the previous set of rules were removed from the terms fields description and narrative and placed in their own dimensions of relevance queries

We added to this topic the filter *timeType="any"* to consider any kind of temporal expression, even unknown ones, in terms of normalization. The places Sumatra and Sri Lanka were found near a question of kind "How long after" so were considered as filters

Other topics like for example topic GeoTime-0014 question about places and dates result in a set of filters, For such topics we created a set of filters including a base filter to remove documents without temporal and geographic references, what is represented with the question mark

Table 10 - Indexed tokens used in filters.

Features	Found values
woeidType	country, city, province
timeType	year, year-month, exact-date, any
place	Yahoo PlaceMaker WOEID references
time	Normalized Expressions found with TIMEXTAG

RUNS

All runs used BM25 Similarity Function:

- INESC-EN-EN-05-DN - Our first run used documents contents index and the base filter to remove documents without geographic or temporal expressions.
- INESC-EN-EN-04-DN - Our second run used sentences index and the base filter to remove sentences without geographic or temporal expressions. The document position was defined by it first sentence in the retrieved list of sentences. Other sentences of that document were ignored.
- INESC-EN-EN-03-DN - Our third run used document content index, the base filter and the filters defined in topic processing, or query expansion if no filters were defined. Bellow we detail this run.
- INESC-EN-EN-02-DN - Our fourth run used the sentences index, a base filter to remove sentences without geographic or temporal expressions and filters defined in topic processing but at sentence level, or query expansion when no filters were defined in the topic.
- INESC-EN-EN-01-DN - Our fifth run used a linear combination of document content and sentences indexes. The base filter, the topic filters and the query

- We also created a complete set of equivalent runs using only the description field in the terms dimension, but keeping the filters that were built using the narrative.

Statistics

Table 1 – Geo-Parsing General Statistics.

	Documents	%
Docs with Places	302695	95,97%
Docs with no Places	12722	0,30%
Docs Failed Annotation	0	0,00%
Docs	315417	100,00

Table 2 – Place types distribution over documents.

WOEID Types	Doc Frequency	References	%References
Town	1047125	1785315	42,75%
Country	419690	965972	23,13%
State	319410	577383	13,82%
POI	210048	307474	7,36%
Suburb	102924	149180	3,57%
County	79251	125312	3,00%
Colloquial	46198	66980	1,60%
Continent	32190	59625	1,43%
Supername	29234	39758	0,95%
ZIP	16604	17122	0,41%
LandFeature	10423	15729	0,38%
Airport	11048	14653	0,35%
Island	9038	12799	0,31%
HistoricalTown	5627	9528	0,23%
Ocean	7052	9475	0,23%
Sea	6321	8443	0,20%
Drainage	4617	6038	0,14%
LocalAdmin	2306	3604	0,09%
Miscellaneous	458	694	0,02%
HistoricalState	477	630	0,02%
Estate	356	460	0,01%
HistoricalCounty	216	317	0,01%
DMA	11	12	0,00%
Market	4	4	0,00%
Zone	2	2	0,00%
Total	2328440	4176509	100,00%

Table 3 – Yahoo Place Maker confidence degree.

Yahoo Conf	Doc Frequency	Refs	%Refs
9	*1071096*	*1989415*	*47,63%*
8	*377001*	*693597*	*16,61%*
10	*314755*	*471296*	*11,28%*
7	*202086*	*354161*	*8,48%*
6	192948	338193	8,10%
5	72404	112156	2,69%
4	52738	81701	1,96%
3	41896	65548	1,57%
2	30541	49241	1,18%
1	12741	21201	0,51%
Total	2368206	4176509	100,00%

Table 4 - Normalized WOEID's.

	Indexed	References
Place WOEID	70477	4176509
Administrative Scopes WOEID	2632	302695
Geographic Scopes WOEID	3752	302695
BelongTos	61299	58640147
All WOEID	138160	63422046

Table 5 – Temporal Expressions general statistics.

	Documents	%
Docs with Timexes	311490	98,75%
Docs with no Timexes found	3809	1,21%
Docs with Indexable Time Exprs	301235	95,50%
Docs with not Indexable Time Exprs	14182	4,50%
Docs Failed Annotation	118	0,04%
All Docs	315417	100,00

Table 6 - Normalized formats statistics.

Expression	Unique tokens	Refs	%
Y	5	229	0,01%
YY	31	11041	0,26%
YYY	80	60734	1,41%
YYYY	3916	18846655	17,44%
YYYY-MM	1297	318580	5,72%
YYYY-	10041	2024089	34,53%
YYYY-Wn	342	100673	2,33%
UNKNOWN	not indexed	1652866	38,31%
Total	15370	4314808	100,00

Table 7 – Duration expressions expanded and indexed.

Expanded Timexes	Direction	Anchor Format	Timexes
PnD (Starting)	STARTING	YYYY-MM-DD	947
PnD (Ending)	ENDING	YYYY-MM-DD	1766
PnW (Starting)	STARTING	YYYY-Wn	1104
PnW (Ending)	ENDING	YYYY-Wn	3936
PnM (Starting)	STARTING	YYYY-MM	1700
PnM (Ending)	ENDING	YYYY-MM	6566
PnY (Starting)	STARTING	YYYY	6786
PnY (Ending)	ENDING	YYYY	50558
Unique Durations Found			5365
References			77781

Table 8 - Duration expressions not used.

Not Used Timexes	Direction	Anchor Format	Timexes
PnD (BEFORE)	BEFORE	YYYY-MM-DD	41880
PnD (AFTER)	AFTER	YYYY-MM-DD	1
PnD (NULL)	NULL	UNKNOWN	271
PnW (BEFORE)	BEFORE	YYYY-Wn	26129
PnW (NULL)	NULL	UNKNOWN	303
PnM (BEFORE)	BEFORE	YYYY-MM	31135
PnM (AFTER)	AFTER	YYYY-MM	1
PnM (NULL)	NULL	UNKNOWN	429
PnY (BEFORE)	BEFORE	YYYY	139020
PnY (AFTER)	AFTER	YYYY	3
PnY (NULL)	NULL	UNKNOWN	1069
Total References			240241

Table 9 – Indexed Temporal Expressions

	Unique	Refs	% to
Key Points	14687	2350436	50,93%
GenPoints	4	104	0,00%
Expanded from Durations	1389	1948788	42,23%
Total - T1	15370	4299328	93,17%
Document DateTime	1363	315417	6,83%
Total - T2 (include doc date)	15370	4614745	100,00

Results

Official results considering Irrelevant, Relevant and Partial Relevant documents:

RUN	MAP	MQ	MNDCG
INESC-EN-EN-01-DN	0.1370	0.1536	0.2961
INESC-EN-EN-02-DN	0.2328	0.2338	0.4056
INESC-EN-EN-03-DN	0.3520	0.3640	0.5641
INESC-EN-EN-04-DN	0.2139	0.2223	0.4234
INESC-EN-EN-05-DN	0.3879	0.4079	0.6246

Results obtained with Treceval considering partially relevant documents as Relevant.

RUN	01-DN	01-D	02-DN	02-D	03-DN	03-D	04-DN	04-D	05-DN	05-D
num_q	25	25	25	25	25	25	25	25	25	25
num_ret	22712	22712	22635	22635	23816	23799	25000	25000	25000	25000
num_rel	1364	1364	1364	1364	1364	1364	1364	1364	1364	1364
num_rel_ret	491	485	608	589	1095	942	686	646	1168	1132
map	0.1523	0.1384	**0.2618**	0.2358	**0.4403**	0.3335	0.2382	0.2320	**0.4213**	**0.3967**
gm_ap	0.0000	0.0000	0.0885	0.0000	0.3009	0.1269	0.0964	0.0000	0.2782	0.2170
ndcg	0.3040	0.2882	**0.3761**	0.3546	**0.6233**	0.5129	0.4048	0.3852	**0.6372**	**0.5989**
R_prec	0.1830	0.1717	0.2794	0.2413	0.4355	0.3236	0.2486	0.2635	0.4191	0.3712
bpref	0.1364	0.1371	0.2571	0.2380	0.3970	0.3227	0.2319	0.2365	0.3810	0.3567
recip_rank	0.1472	0.1790	0.1301	0.1484	0.1563	0.1173	0.1511	0.1675	0.2545	0.1812
P5	0.2640	0.2800	**0.4960**	0.4320	**0.6560**	0.5200	**0.4800**	0.4720	**0.6000**	**0.5520**
P10	0.2160	0.1960	0.3600	0.3160	0.5560	0.4240	0.3440	0.3320	0.5240	0.4880
P15	0.1813	0.1680	0.3067	0.2800	0.4987	0.3547	0.3147	0.2827	0.4613	0.4533
P20	0.1540	0.1640	0.2680	0.2440	0.4440	0.3340	0.2800	0.2680	0.4200	0.4300
P30	0.1347	0.1333	0.2320	0.2120	0.3867	0.3000	0.2267	0.2173	0.3880	0.3760
P100	0.0896	0.0860	0.1496	0.1228	0.2324	0.1844	0.1488	0.1392	0.2428	0.2332
P200	0.0588	0.0592	0.0938	0.0840	0.1474	0.1234	0.0950	0.0882	0.1624	0.1520
P500	0.0327	0.0317	0.0442	0.0430	0.0761	0.0646	0.0486	0.0470	0.0806	0.0790
P1000	0.0196	0.0194	0.0243	0.0236	0.0438	0.0377	0.0274	0.0258	0.0467	0.0453

inesc id lisboa
INSTITUTO SUPERIOR TÉCNICO
ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO
INSTITUTO POLITÉCNICO de PORTALEGRE