Using various indexing schemes and multiple translations in the CL-SR task at CLEF 2005

Diana Inkpen, Muath Alzghool, and Aminul Islam
University of Ottawa
{diana,alzghool,mdislam}@site.uottawa.ca

System overview

- Smart IR system
- Online MT tools

- Task
  - Collection – ASR transcribed text
  - Training queries (38), test queries (25)
  - Relevance judgments

Results of the five submitted runs, for topics in English, Spanish, French, and German

Language	Run	map	bpref	Fields	Description
English	uoEnTDN	0.2176	0.2005	TDN	Weighting scheme: mpc/ntn
Spanish	uoSpTDN	0.1863	0.1750	TDN	Weighting scheme: mpc/ntn
French	uoFrTD	0.1685	0.1599	TD	Weighting scheme: mpc/ntn
English	**uoEnTD**	**0.1653**	**0.1705**	**TD**	**Weighting scheme: mpc/ntn**
German	uoGrTDN	0.1281	0.1331	TDN	Weighting scheme: mpc/ntn

Translating queries with online MT tools

Spanish, German, French:

1. http://www.google.com/language_tools?hl=en
2. http://www.babelfish.altavista.com
3. http://freetranslation.com
4. http://www.wordlingo.com/en/products_services/wordlingo_translator.html
5. http://www.systranet.com/systran/net
6. http://www.online-translator.com/srvurl.asp?lang=en
7. http://www.freetranslation.paralink.com

Czech:

1. http://intertran.tranexp.com/Translate/result.shtml

Example query

<top>
<num>1159
<title>Child survivors in Sweden
<desc>Describe survival mechanisms of children born in 1930-1933 who spend the war in concentration camps or in hiding and who presently live in Sweden.
<narr>The relevant material should describe the circumstances and inner resources of the surviving children. The relevant material also describes how the wartime experience affected their post-war adult life.
</top>

<top>
<num>1159
<title>Les enfants survivants en Suède
<desc>Descriptions des mécanismes de survie des enfants nés entre 1930 et 1933 qui ont passé la guerre en camps de concentration ou cachés et qui vivent actuellement en Suède.
</top>

Example of translated query (from French)

<top>
<num> 1159
<title> surviving children in Sweden
surviving children in Sweden
The children survivors in Sweden
surviving children in Sweden
surviving children in Sweden
The surviving children in Sweden
surviving children in Sweden
<desc> Descriptions of the mechanisms of survival of the children born between 1930 and 1933 who passed the war in concentration camps or hidden and who currently live in Sweden.

Descriptions of the mechanisms of survival of the children born between 1930 and 1933 who passed the war in concentration camps or hidden and who currently live in Sweden.

Descriptions of the survival mechanisms of the born children between 1930 and 1933 that passed the war in concentration camps or hidden and that live currently in Sweden.

Descriptions of the mechanisms of survival of the children born between 1930 and 1933 who passed the war in concentration camps or hidden and who currently live in Sweden. … …

<narr>
</top>

Results on the output of each Machine Translation system: Spanish, French, German, and Czech

Measure	Translation							
	Sp1	**Sp2**	**Sp3**	**Sp4**	**Sp5**	**Sp6**	**Sp7**	**Spanish All**
map	0.1711	0.1756	0.1758	0.1563	0.1756	0.1784	0.1756	0.1863
bpref	0.1708	0.1733	0.1637	0.1563	0.1733	0.1739	0.1733	0.1750
	Fr1	**Fr2**	**Fr3**	**Fr4**	**Fr5**	**Fr6**	**Fr7**	**French All**
map	0.1547	0.1551	0.1526	0.1562	0.1551	0.1575	0.1551	0.1685
bpref	0.1554	0.1559	0.1551	0.1572	0.1559	0.1668	0.1559	0.1599
	Gr1	**Gr2**	**Gr3**	**Gr4**	**Gr5**	**Gr6**	**Gr7**	**German All**
map	0.1244	0.1238	0.1189	0.1232	0.1239	0.1491	0.1238	0.1281
bpref	0.1281	0.1286	0.1344	0.1279	0.1287	0.1633	0.1287	0.1331
	Czech							
map	0.1166							
bpref	0.1310							

Smart IR system

- (Buckley et. *al*., 2000),
- Stemming, pseudo-relevance.

- Many weighting schemes (60 x 60)
  - For document terms
  - For query terms

Weighting schemes for documents and queries

- **Term frequency component**

none (n) : $new\_tf = tf$

max-norm (m) : $new\_tf = \frac{tf}{\max\_tf}$

augmented normalized (a): $new\_tf = 0.5 + 0.5 * (\frac{tf}{\max\_tf})$

log (l): $new\_tf = \ln(tf) + 1.0$

square (s): $new\_tf = tf^2$

- **Merging of collection frequency component**

none (n): $new\_wt = new\_tf$

inverse document frequency weight (t): $new\_wt = new\_tf * \log(\frac{num\_docs}{coll\_freq\_of\_term})$

probabilistic (p): $new\_wt = new\_tf * \log(\frac{num\_docs - coll\_freq}{coll\_freq})$

squared (s): …

- **Merging of vector normalization**

none (n): $norm\_wt = new\_wt$

sum (s): $norm\_wt = \frac{tf}{\sum_m new\_wt}$

cosine (c): $norm\_wt = \frac{tf}{\sqrt{\sum_m new\_wt^2}}$

Results of the various indexing (weighting) schemes, for English topics

	Weighting scheme	TDN		TD		T	
		map	bpref	map	bpref	map	bpref
1	mpc/mts	0.2175	0.2004	0.1651	0.1707	0.1175	0.1374
2	mpc/nts	0.2175	0.2004	0.1651	0.1707	0.1175	0.1374
3	mpc/ntn	**0.2176**	0.2005	0.1653	0.1705	0.1174	0.1371
4	npc/ntn	**0.2176**	0.2005	0.1653	0.1705	0.1174	0.1371
5	mpc/mtc	**0.2176**	0.2005	0.1653	0.1705	0.1174	0.1371
6	mpc/ntc	**0.2176**	0.2005	0.1653	0.1705	0.1174	0.1371
7	mpc/mtn	**0.2176**	0.2005	0.1653	0.1705	0.1174	0.1371
8	npn/ntn	0.2116	0.1916	**0.1681**	0.1693	0.1181	0.1350
9	lsn/ntn	0.1195	0.1487	0.1233	0.1433	**0.1227**	0.1395
10	lsn/atn	0.0919	0.1456	0.1115	0.1355	**0.1227**	0.1395
11	asn/ntn	0.0912	0.1295	0.0923	0.1208	0.1062	0.1290
12	snn/ntn	0.0693	0.1327	0.0592	0.1305	0.0729	0.1113
13	sps/ntn	0.0349	0.0979	0.0377	0.1036	0.0383	0.0783
14	nps/ntn	0.0517	0.0940	0.0416	0.0791	0.0474	0.0761
15	mtc/atc	0.1138	0.1514	0.1151	0.1449	0.1108	0.1345

Phonetic transcripts

- The documents and the queries were transcribed in phonetic form and split into 4-grams.

- Example:

<top>

<num> 1159

<title> ch_ay_l_d s_ax_r_v ax_r_v_ay r_v_ay_v v_ay_v_ax ay_v_ax_r v_ax_r_z ih_n s_w_iy_d w_iy_d_ax iy_d_ax_n

<desc>

<narr>

</top>

Results on phonetic n-grams, and combination text plus phonetic n-grams

Language	map	bpref	Fields	Description
English	0.1276	0.1117	T	Phonetic, mpc/ntn
English	0.2550	0.1492	TD	Phonetic, mpc/ntn
English	0.1245	0.1198	T	Phonetic+Text, mpc/ntn
English	0.2590	0.1585	TD	Phonetic+Text, mpc/ntn
Spanish	0.1395	0.1050	T	Phonetic, mpc/ntn
Spanish	0.2653	0.1549	TD	Phonetic, mpc/ntn
Spanish	0.1443	0.1108	T	Phonetic+Text, mpc/ntn
Spanish	0.2669	0.1576	TD	Phonetic+Text, mpc/ntn
French	0.1251	0.1005	T	Phonetic, mpc/ntn
French	0.2726	0.1747	TD	Phonetic, mpc/ntn
French	0.1254	0.1023	T	Phonetic+Text, mpc/ntn
French	0.2833	0.1841	TD	Phonetic+Text, mpc/ntn
German	0.1163	0.1150	T	Phonetic, mpc/ntn
German	0.2356	0.1568	TD	Phonetic, mpc/ntn
German	0.1187	0.1159	T	Phonetic+Text, mpc/ntn
German	0.2324	0.1601	TD	Phonetic+Text, mpc/ntn

Results of indexing all fields: manual keywords and summaries, ASR transcripts

Language	map	bpref	Fields	Description
English	**0.4647**	0.3660	TDN	Weighting scheme: mpc/ntn, Manual fields
Spanish	0.3811	0.2988	TDN	Weighting scheme: mpc/ntn, Manual fields
French	0.3496	0.2864	TD	Weighting scheme: mpc/ntn, Manual fields
German	0.2513	0.2656	TDN	Weighting scheme: mpc/ntn, Manual fields
Czech	0.2338	0.2251	TDN	Weighting scheme: mpc/ntn, Manual fields