ADVANCED FUZZY MATCHING IN THE TRANSLATION OF EU TEXTS

Margita Šoštarić

Appendices

Appendix I.

Matching algorithms and configurations

Formulae for the calculation of similarity between the query (Q) and the source side of the match (S), or the query and its reference translation (R) in case of MT evaluation metrics. The formulae which are not given here can be found in the references given in thesis.

Levenshtein distance:

$$LEV(Q,S_i) = 1 - (\Delta_{LEV}(Q,S_i) / \max(|Q|,|S_i|))$$

Percent match:

$$PM(Q,S_i) = |Q_{unigrams}| \cap |S_{i,unigrams}| / |Q_{unigrams}|$$

Ngram precision:

$$NGP = \Sigma_{n=1}^{N} (|Q_{n\text{-}grams} \cap S_{i,n\text{-}grams}|) / (Z * |Q_{n\text{-}grams}| \mid (1\text{-}Z) * |S_{i,n\text{-}grams}|)/N$$

Normalized TER:

$$TER(Q,R) = 1 - (\log(1+\Delta_{TER}(Q,R) / |R|))/3)$$

Table 1: All filtering, matching and evaluation configurations.

METRIC	MATCH ITEM	WEIGHTING SCHEME	SPECIAL PARAMETERS
Approximate query coverage	word part sequences, 1	default	threshold: 0.2 nbest: 50
Baseline (Levenshtein)	words, 1	default	/
BEER	words, 1	default	all modules for EN
Levenshtein	lemmas, 1	default	/
Levenshtein	lemmas, 6	ignore case	/
Levenshtein	POS-tags, 4	default	/
Levenshtein	Prüfer sequences, 4	default	/
METEOR	words, 1	default	all modules for EN
Ngram precision	words, 1	default	N = 4, Z =0.3
Ngram precision	lemmas, 4	default	N = 4, Z =0.3
Ngram precision	POS-tags, 4	default	N = 4, Z =0.3
Ngram precision	Prufer sequences, 4	Prüfer weights	N = 4, Z =0.3
Percent match	words, 1	default	/
Percent match	lemmas, 3	default	/
Percent match	POS-tags, 3	default	/
Percent match	Prüfer sequences, 2	default	/
Shared partial subtrees	parse	default	/
TER	words, 1	default	/
$METEOR_T$	words, 1	default	exact, stem, paraphrase modules for SE
$Ngram\ precision_T$	words, 2:1	default	N=4, Z =0
$Shared\ partial\ subtrees_T$	parse	default	/
TER_T	words, 1	default	/

Appendix II.

An example from the main group of survey questions.

***The agenda shall be adopted by the Trade Committee at the beginning of each meeting.**

This question is mandatory.

- ○ < The > budget < shall be adopted by > < the > Commission.

  Budgeten ska antas av kommissionen.
- ◉ < The > final < agenda shall be adopted > < at the beginning of each meeting. >

  Den slutliga dagordningen skall antas i början av varje sammanträde.
- ○ Both equal

Appendix III.

Automatic evaluation of the data

Table 2: Automatic evaluation for the range above or equal to 70 percent overlap.

	MET_T corr	MET_T mean	NGP_T corr	NGP_T mean	SPS_T corr	SPS_T mean	TER_T corr	TER_T mean
BASELINE	0.4491	0.7516	0.4373	0.6574	0.4527	0.7554	0.4702	0.2342
BEER	**0.5005**	**0.7585**	**0.5036**	**0.6663**	0.4634	**0.7590**	0.4223	0.2341
$LEV_{LEM1DEF}$	0.2948	0.7442	0.2569	0.6465	0.3475	0.7514	0.4006	0.2392
$LEV_{LEM6IGN}$	0.2801	0.7447	0.2541	0.6481	0.3076	0.7507	0.3349	0.2411
$LEV_{POS4DEF}$	0.0737	0.7320	0.0479	0.6325	0.1242	0.7372	0.2103	0.2518
$LEV_{PRUF4DEF}$	0.1934	0.7484	0.1837	0.6524	0.2477	0.7512	0.2581	0.2380
METEOR	0.3564	0.7555	0.3666	0.6637	0.3442	0.7527	0.3123	0.2390
$NGP_{WORD1DEF}$	0.473	0.7552	0.4913	0.6636	0.3701	0.7543	0.3153	0.2386
$NGP_{LEM4DEF}$	0.3131	0.7438	0.2909	0.6475	0.3106	0.7480	0.3132	0.2456
$NGP_{POS4DEF}$	0.0897	0.7320	0.0691	0.6325	0.1275	0.7357	0.1980	0.2544
$NGP_{PRUF4PRUF}$	0.2442	0.7486	0.2399	0.6540	0.2714	0.7502	0.2550	0.2415
$PM_{WORD1DEF}$	0.3556	0.7433	0.3455	0.6469	0.3344	0.7463	0.3378	0.2425
$PM_{LEM3DEF}$	0.2966	0.7418	0.2654	0.6437	0.3164	0.7482	0.3436	0.2434
$PM_{POS3DEF}$	0.0657	0.7299	0.0380	0.6298	0.1118	0.7356	0.1960	0.2542
$PM_{PRUF2DEF}$	0.1536	0.7442	0.1432	0.6469	0.2114	0.7485	0.2240	0.2415
SPS	0.2198	0.7474	0.1962	0.6513	0.2840	0.7537	0.2913	0.2370
TER	0.4439	0.7521	0.4323	0.6574	**0.4643**	0.7562	**0.4913**	**0.2324**
ALL	0.4428	0.7523	0.4541	0.6589	0.4058	0.7557	0.4289	0.2344

Table 3: Automatic evaluation for the range below 70 percent overlap.

	MET_T corr	MET_T mean	NGP_T corr	NGP_T mean	SPS_T corr	SPS_T mean	TER_T corr	TER_T mean
BASELINE	0.6735	0.3372	0.5928	0.2502	0.6667	0.3640	0.6162	0.8687
BEER	0.6328	0.3499*	0.6031	0.2650*	0.6002	0.3650	0.4198	0.9962
$LEV_{LEM1DEF}$	0.6477	0.3347	0.5559	0.2466	0.6553	0.3643	0.6202	0.8681
$LEV_{LEM6IGN}$	0.6396	0.3197	0.5905	0.2438	0.6182	0.3493	0.5279	0.8900
$LEV_{POS4DEF}$	0.4930	0.3112	0.4238	0.2309	0.5059	0.3431	0.5032	0.8970
$LEV_{PRUF4DEF}$	0.5202	0.3136	0.4676	0.2332	0.5351	0.3435	0.4709	0.9096
METEOR	0.6967	0.3534*	0.6723	0.2752*	0.5601	0.3531	0.3619	1.0049
$NGP_{WORD1DEF}$	0.7277	**0.3546***	0.7004	**0.2764***	0.6087	0.3576	0.3587	1.0018
$NGP_{LEM4DEF}$	0.6732	0.3358	0.6402	0.2614*	0.5902	0.3517	0.3935	0.9530
$NGP_{POS4DEF}$	0.5385	0.3158	0.4863	0.2383	0.5067	0.3398	0.4049	0.9453
$NGP_{PRUF4PRUF}$	0.5585	0.3202	0.5334	0.2432	0.5183	0.3412	0.3306	0.9635
$PM_{WORD1DEF}$	0.5577	0.3111	0.4862	0.2204	0.5116	0.3494	0.5956	0.7750*
$PM_{LEM3DEF}$	0.6175	0.3220	0.5497	0.2379	0.5869	0.3534	0.5934	0.8259*
$PM_{POS3DEF}$	0.4872	0.3008	0.4237	0.2149	0.4756	0.3410	0.5540	0.8084*
$PM_{PRUF2DEF}$	0.4738	0.3018	0.4135	0.2142	0.4736	0.3436	0.5358	0.7930*
SPS	0.5625	0.3196	0.4919	0.2281	0.5804	0.3653	0.5415	0.8095*
TER	0.6533	0.3155	0.5841	0.2235	0.6607	0.3666	**0.7075**	**0.7443***
ALL	**0.7663**	0.3524*	**0.7255**	0.2687*	**0.7089**	**0.3705***	0.6566	0.8395*

Appendix IV.

Pearson correlations on the human-evaluated subset

Table 4: Pearson correlations between the fuzzy matches and the human and automatic evaluation. Results which are higher than the baseline are bolded, statistically insignificant results (p>0.05) are in italics.

	HUM corr	MET_T corr	NGP_T corr	SPS_T corr	TER_T corr
BASELINE	0.2087	0.3637	0.3189	0.4092	0.3731
BEER	**0.2407**	**0.5565**	**0.4836**	**0.5116**	0.2748
$LEV_{LEM1DEF}$	**0.2182**	**0.3714**	**0.3220**	**0.4170**	**0.3777**
$LEV_{LEM6IGN}$	**0.2377**	**0.4820**	**0.4963**	**0.4543**	0.3233
$LEV_{POS4DEF}$	0.1816	0.2435	0.2729	0.2676	0.2010
$LEV_{PRUF4DEF}$	0.1464	0.1274	*0.1416*	0.1612	*0.1120*
METEOR	0.1996	**0.6293**	**0.5579**	**0.4987**	0.2101
$NGP_{WORD1DEF}$	**0.2164**	**0.6331**	**0.5946**	**0.5054**	0.2623
$NGP_{LEM4DEF}$	**0.2466**	**0.5986**	**0.6043**	**0.4950**	0.2880
$NGP_{POS4DEF}$	0.1938	0.3547	**0.3874**	0.3106	0.1566
$NGP_{PRUF4PRUF}$	0.1575	0.2732	0.2787	0.2246	*0.0937*
$PM_{WORD1DEF}$	**0.3373**	0.3464	0.2681	0.3895	**0.4518**
$PM_{LEM3DEF}$	**0.3100**	**0.5383**	**0.5044**	**0.5221**	**0.4427**
$PM_{POS3DEF}$	**0.2403**	0.2189	0.2121	0.2368	0.2691
$PM_{PRUF2DEF}$	**0.2270**	*0.0519*	*0.0297*	*0.1209*	0.1822
SPS	0.1955	0.2640	0.2121	0.3403	0.2481
TER	*0.1146*	0.1568	*0.0899*	0.3515	**0.5412**
ALL	0.2407	0.3789	0.3369	0.3383	0.2028