Machine Translation of Bio-Thesauri

Erik M. van Mulligen[1], Quoc-Chinh Bui[1], Jan A. Kors[1]

[1]Department of Medical Informatics, Erasmus University Medical Center Rotterdam, The Netherlands
{e.vanmulligen,q.bui,j.kors}@erasmusmc.nl

Abstract. In this paper we describe how we applied a general- purpose machine translation tool for translating biomedical thesauri. We used corresponding terms in parallel corpora to check the validity of the translations. The advantage of this approach is that a single corresponding set of terms can be verified where techniques to retrieve translations from a parallel corpus do not exploit the knowledge contained in current state of the art machine translation software.

Keywords: machine translation, multilingual, concept annotation

1 Introduction

The Unified Medical Language System [1] shows that a variety of thesauri can be integrated in a single system, the MetaThesaurus. The majority of these thesauri are in English and only a few of them contain translations in other languages. If these translated thesauri are used for concept normalization tasks it is evident that the performance will be lower than for English.

In the MANTRA project [2] we investigate possibilities to automatically enrich the translations of English thesauri. One possibility is to mine parallel corpora for associations between terms in English and other languages[3,4]. The disadvantage of this approach is that one needs multiple associations before one can infer a translation [5]. In this paper we describe an alternative approach where we use a general machine translation service to translate the thesaurus terms into another language. Subsequently we verify the quality of the translations by checking in a parallel corpus whether we find for a term in an English sentence the translated term in the corresponding non-English sentence. The advantage is that also associations that occur only once can be found and used as a proper term translation. Furthermore, this approach yields also translations for terms that have no association in the parallel corpus. These can be verified against the terms that are available in the non-English thesaurus. Finally, translated terms can also be manually verified [6].

2 Methods

For the thesaurus translation we used Google Translate[7]. The English thesaurus in our experiments was the thesaurus created in the MANTRA project. The MANTRA thesaurus is a subset from the 2012AB UMLS resources. It contains all terms belonging to concepts in MeSH, MEDDRA, and SNOMED-CT, but excludes the semantic types in the semantic groups Activities, Concepts, Genes, Occupations, Organisms, and Procedures[8]. In addition, some terms that have particular term types are removed from the MANTRA thesaurus.

All terms from the MANTRA thesaurus where fed to the API of Google Translate. The response included one or more candidate translations for each term with a score for each candidate. We took the candidate term with the highest score and concatenated the individual translated sections - Google provides basically translations per word - into a translated term. Non-English terms already contained in the UMLS were also included in the translated thesaurus. We also created a second translated thesaurus only based on the existing non-English terms in the UMLS. This UMLS-translated thesaurus was included as a baseline thesaurus to assess the improvement that could be obtained with the machine translation. We applied this approach for two languages, Dutch and German.

We used two parallel corpora: multilingual drug labels from the EMEA corpus, and bilingual titles of scientific abstracts from Medline.

3 Results

We indexed the parallel corpora with the English thesaurus, the UMLS-translated thesaurus, and the machine-translated thesaurus. We differentiated for the different semantic types and the different parallel corpora we have (for German restricted to the drugs labels from EMEA and MedLine titles The results for Dutch, German, and French are show in table 1. Each table shows the results for finding concepts using only the manual translation as contained in the UMLS and the results when the machine translated terms are added. We provide not only the figures for the translated terms that have correspondences in English and the translation language (BOTH), but also terms that have only be found in English (ENGLISH) or only in the translation language (DUTCH, GERMAN, or FRENCH). The tables show the results for the EMEA drug label parallel corpus and for the Medline titles.

Tabel 1. Results for the Dutch, German, and French EMEA drug labels and Medline titles.

	ENTRY	OBJC	GEOG	CHEM	DEVI	PHEN	DISO	ANAT	LIVB	PHYS	TOTAL
Dutch EMEA original	DUTCH	28	7	8	13	19	414	68	81	57	695
	ENGLISH	356	125	4115	182	222	3215	814	570	467	10066
	BOTH	58	82	17	19	71	2231	336	275	217	3306
EMEA machine translation	DUTCH	120	15	1037	74	59	1115	303	176	140	3039
	ENGLISH	193	71	1014	91	105	1793	421	314	205	4207
	BOTH	221	136	3118	110	188	3653	729	531	479	9165
Medline original	DUTCH	41	8	5	10	18	541	73	27	174	897
	ENGLISH	1287	378	18381	956	801	21892	6537	1719	7069	59020
	BOTH	103	137	12	68	72	2873	445	196	415	4321
Medline machine translation	DUTCH	151	24	732	78	52	1160	317	112	293	2919
	ENGLISH	1100	346	16719	873	728	20642	6135	1523	6623	54689
	BOTH	290	169	1674	151	145	4123	847	392	861	8652
German EMEA original	GERMAN	45	7	143	59	28	445	74	63	82	946
	ENGLISH	329	84	2866	156	222	3438	784	527	442	8848
	BOTH	85	123	1266	45	71	2008	366	318	242	4524
Medline machine translation	GERMAN	127	21	1093	73	72	1209	338	189	179	3301
	ENGLISH	174	19	1026	91	127	2019	429	257	214	4356
	BOTH	240	188	3106	110	166	3427	721	588	470	9016
Medline original	GERMAN	67	19	262	157	56	1146	97	71	327	2202
	ENGLISH	1118	262	14785	823	647	18799	5825	1290	5638	49187
	BOTH	272	253	3608	201	226	5966	1157	625	1846	14154
Medline machine translation	GERMAN	252	36	1777	141	123	2208	634	171	546	5888
	ENGLISH	711	186	9068	581	468	15074	4102	910	3640	34740
	BOTH	679	329	9325	443	405	9691	2880	1005	3844	28601

French **EMEA original**	FRENCH	42	16	217	11	43	590	94	84	78	1175
	ENGLISH	254	79	2317	157	180	2496	607	381	353	6824
	BOTH	87	102	1289	23	87	2204	380	298	268	4738
EMEA machine translation	FRENCH	157	28	1198	85	78	1382	326	244	173	3671
	ENGLISH	109	23	717	66	82	1590	307	155	176	3225
	BOTH	232	158	2889	114	185	3110	680	524	445	8337
Medline original	FRENCH	71	22	430	38	69	1747	159	152	334	3022
	ENGLISH	712	128	7726	554	377	10527	3773	798	2745	27340
	BOTH	290	298	3566	150	282	7360	1315	728	2102	16091
Medline machine translation	FRENCH	283	42	1999	182	143	3321	873	272	726	7841
	ENGLISH	284	48	3133	277	199	7022	2125	398	874	14360
	BOTH	718	378	8159	427	460	10865	2963	1128	3973	29071

Tabel 2. Overall statistics for the different languages and corpora.

	EMEA		Medline	
Language	**Original**	**Translated**	**Original**	**Translated**
Dutch	3039	9165	4321	8652
German	4524	8921	14154	28601
French	4738	8337	16091	29071

4 Discussion

The results show that machine translation can help to enrich a thesaurus. Compared with the manual UMLS-translated thesaurus, the number of terms in the machine-translated thesaurus that are found in the parallel corpora, doubles for German, French and almost triples for Dutch when only considering concepts that have been found in English. This is consistent for both parallel corpora included in this evaluation. The German and French manual translations are more extensive than the Dutch one, which likely explains the difference in number of extra terms found. The increase is largest for some semantic groups that have hardly been translated (objects, devices, and chemicals). If one also looks at terms that have only been found in the translated corpus and not in the original English corpus an additional set of translated terms can be found. We will also evaluate this set of terms only found in the translated corpus for correctness. We will extend this evaluation to include Spanish as well.

References

1. Lindberg DA, Humphreys BL, McCray AT . The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91.
2. http://mantra-project.eu
3. Deleger L, Merkel M, Zweigenbaum P. Translating medical terminologies through word alignment in parallel text corpora. Journal of Biomedical Informatics 2009; 42:692-701.

4. van der Plas L, Tiedemann J. Finding medical term variations using parallel corpora and distributional similarity. In: Proc 6th Workshop on Ontologies and Lexical Resources, 2010.
5. Och FJ, Ney H. A systematic comparison of various statistical alignment models. Comp Ling 2003;29:19-51.
6. Schulz S. et al. Machine vs. human translation of SNOMED CT terms. Currently under review for MEDINFO 2013
7. https://developers.google.com/translate/
8. Bodenreider and O, McCray, AT. Exploring semantic groups through visual approaches, Journal of Biomedical Informatics 36(6):414–432, 2003.