Comparable Fora

Johanka Spoustová **Miroslav Spousta**
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics,
Charles University Prague, Czech Republic
{johanka,spousta}@ufal.mff.cuni.cz

Abstract

As the title suggests, our paper deals with web discussion fora, whose content can be considered to be a special type of comparable corpora. We discuss the potential of this vast amount of data available now on the World Wide Web nearly for every language, regarding both general and common topics as well as the most obscure and specific ones. To illustrate our ideas, we propose a case study of seven wedding discussion fora in five languages.

1 Introduction to comparable corpora

Nearly every description of comparable corpora begins with the EAGLES (Expert Advisory Group on Language Engineering Standards) definition:[1]

"A comparable corpus is one which selects similar texts in more than one language or variety. The possibilities of a comparable corpus are to compare different languages or varieties in similar circumstances of communication, but avoiding the inevitable distortion introduced by the translations of a parallel corpus."

(Maia, 2003), which also became nearly standard during the recent years, emphasizes the fact that comparable monolingual corpora usually provide us with much better linguistic quality and representativeness than translated parallel corpora. The other advantages over the parallel corpora, i.e. amount and availability, are obvious.

Nowadays, the most popular usage of comparable corpora is improving machine translation, more precisely, compensating the lack of parallel training data. The articles (Munteanu et al., 2004), (Munteanu and Marcu, 2005) and (Munteanu and Marcu, 2006) are introducing algorithms for extracting parallel sentences and sub-sententional fragments from comparable corpora and using the automatically extracted parallel data for improving statistical machine translation algorithms performance.

Present day most popular comparable corpora come either from the newswire resources (AFP, Reuters, Xinhua), leading to data sets like LDC English, Chinese and Arabic Gigaword, or from Wikipedia. Mining Wikipedia became very popular in the recent years. For example, (Tomás et al., 2008) is exploring both parallel and comparable potential of Wikipedia, (Filatova, 2009) examines multilingual aspects of a selected subset of Wikipedia and (Gamallo and López, 2010) describes converting Wikipedia into "CorpusPedia".

2 Introduction to fora

Just to avoid confusion: In this article, we focus only on fora or boards, i.e. standalone discussion sites on a stated topic. We are not talking about comments accompanying news articles or blog posts.

The internet discussion fora cover, in surprisingly big amounts of data and for many languages, the most unbelievable topics (real examples from the authors' country). People, who eat only uncooked ("raw") food. People, who eat only cooked food. Mothers with young children, women trying to conceive, communities of people absolutely avoiding sex. Fans of Volvo, BMW, Maserati, Trabant cars. Probably also in your country mothers like to talk

[1] http://www.ilc.cnr.it/EAGLES96/corpustyp/node21.html

Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 96–101,
49th Annual Meeting of the Association for Computational Linguistics,
Portland, Oregon, 24 June 2011.

about their children and men like to compare their engine's horse power.

Everyone who has any specific interest or hobby and is friendly with the web, probably knows at least one discussion forum focused on his/her favourite topic, inhabited by intelligent, friendly debaters producing interesting, on-topic content. These types of fora often have very active administrators, who clean the discussions from off-topics, vulgarities, move the discussion threads into correct thematic categories etc. The administrators' "tidying up" effort can be even regarded as a kind of annotation.

The rapidly growing amount of web discussion fora was until now linguistically exploited only in a strictly monolingual manner. To the best of our (and Google Scholar) knowledge, nobody has published any work regarding the possibility of using internet discussion fora as a multilingual source of data for linguistic or machine translation purposes.

2.1 Forum structure

A typical forum is divided into thematic categories (larger fora split into boards and boards into categories). Every category usually contains from tens to thousands of separate discussions. A discussion consists of messages (posts) and sometimes its content is further arranged using threads.

A discussion should be placed in appropriate category and messages in the discussion should hold onto the discussion topic, otherwise the administrator removes the inappropriate messages or even the whole discussion.

Fora usually have an entire off-topic category where their members can talk about anything "out-of-domain".

To avoid spam, usually only registered members can contribute. Some fora keep their memberlist visible to the public, some do not.

3 Why comparable fora?

Besides their amount and availability, comparable fora have a few other advantages over other types of comparable corpora.

They contain "spontaneous writing" – an original, previously unpublished content, which is almost certainly not a translation of other language original. This is obviously not the case of parallel corpora, and we cannot be sure even for other popular comparable corpora. A journalist may be inspired by a news agency report or by another media source, and a Wikipedia author must also reconcile his claims with existing resources, which more or less affects his writing style.

The other advantage is easier domain classification, or more effective pre-selection before running an automatic parallel sentences alignment. A generic newspaper article is provided only with a title, language and release date. A Wikipedia entry has a title, history and is classified into a thematic category. Fora messages have both dates, titles and category classifications and they are available in much larger amounts than Wikipedia entries and are covering more thematic domains than news articles.

4 A case study: wedding sites

As a topic of our case study, we have chosen an event which occurs to most of the people at least once in their life – a wedding.

4.1 General overview

We looked over five language mutations of the same forum operated by Asmira Company – Finalstitch.co.uk (EN), Braupunkt.de (DE), Fairelanoce.fr (FR), Mojasvadba.sk (SK), Beremese.cz (CZ); and two other fora, Brides.com/forums (EN2) and Organisation-mariage.net (FR2), which seem to be larger and more popular in the target countries.

We have manually examined fora sizes and possibilities of their alignment on the category level.

Tables 1 and 2 summarize the total number of discussions and messages contained in selected categories, shared by most of the fora. For the Asmira fora, we omitted the discussions accessible both from CZ and SK sites.

If we assume average length of a message to be about 60 words (see below), the proposed sites give us a few millions of words of multilingual comparable corpora in each category (focussed on very restricted topic, such as wedding clothes, or hairdressing & make-up) even for "non-mainstream" languages, such as Czech or Slovak.

4.2 Quantitative characteristics

In order to learn more about the amount and textual quality of the data, we have downloaded all the con-

	EN	DE	FR	CZ	SK	EN2	FR2
Ceremony and reception	389	280	232	1 532	2 345	N/A	1 536
Wedding-preparations	474	417	654	916	1270	13632	1 873
Date & location	63	119	154	839	529	371	N/A
Beauty	68	47	74	472	794	2 858	2 452
Wedding clothing	291	166	200	715	1 108	10 832	
After the wedding	37	47	47	236	245	1 530	390

Table 1: Total number of discussions in the selected wedding fora.

	EN	DE	FR	CZ	SK	EN2	FR2
Ceremony and reception	3 863	3 947	4 174	43 436	64 273	N/A	19 002
Wedding-preparations	4 908	4 987	8 867	51 880	27 837	130 408	24 585
Date & location	1 004	1 988	3 178	550 969	279 091	24 513	N/A
Beauty	692	852	1 462	32 118	32 620	15 946	38 582
Wedding clothing	2 634	2 336	3 588	27 624	28 048	75 331	
After the wedding	527	1 012	1 065	30 588	18 090	23 612	6 286

Table 2: Total number of messages in the selected wedding fora.

tent of the five Asmira fora, extracted their messages into five monolingual corpora and measured some basic characteristics of the texts. The downloading and extracting task needed about 20 minutes of coding and a few days of waiting for the result (we did not want to overload the fora webservers).

Table 3 shows us average messages lengths (in words) for particular categories of these fora.

In graphs 1, 2 and 3, we present normalized sentence length distributions for particular fora. For English and Czech, we added for comparison sentence length distributions of reference corpora of comparable sizes, i.e. The Penn Treebank, training set (Marcus et al., 1994), for English and The Czech National Corpus, SYN2005 (CNC, 2005), for Czech.

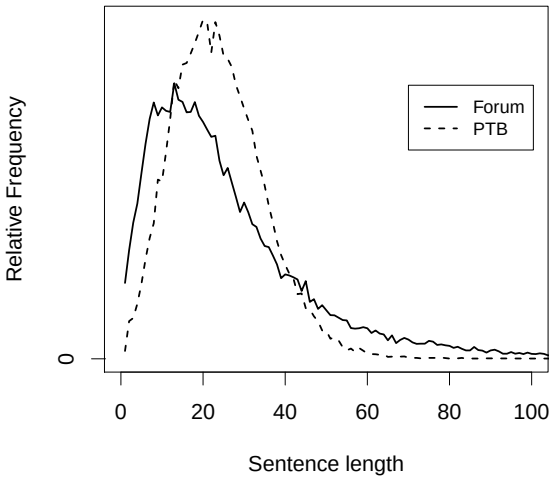

Figure 1: The EN forum and The Penn Treebank - sentence length distributions.

4.3 Examples of similar discussion topics

The category distinction may be still too coarse for potential alignment. The site FR2 has a joint category for Beauty and Wedding clothing, and on the contrary, it has separate categories for Wedding and Reception. Therefore, we tried to examine the fora on a deeper level. In table 4, we present some examples of discussions on the same topic.

As you can guess, fully automatic alignment of the discussion titles will not be an easy task. On the other side, every machine translation specialist must shiver with pleasure when seeing some of the discussion titles to be almost translations of each other, and it would be a sin to leave these data unexploited.

	EN	DE	FR	CZ	SK
Ceremony and reception	70.0	68.7	51.9	59.7	56.9
Wedding-preparations	73.8	62.5	55.1	63.7	62.3
Date & location	59.2	56.4	61.7	52.0	48.8
Beauty	67.7	61.3	53.4	65.8	56.6
Wedding clothing	61.1	60.4	42.1	57.0	50.0
After the wedding	71.8	69.5	52.0	66.8	68.6

Table 3: Average messages lengths (in words) for the selected wedding fora categories.

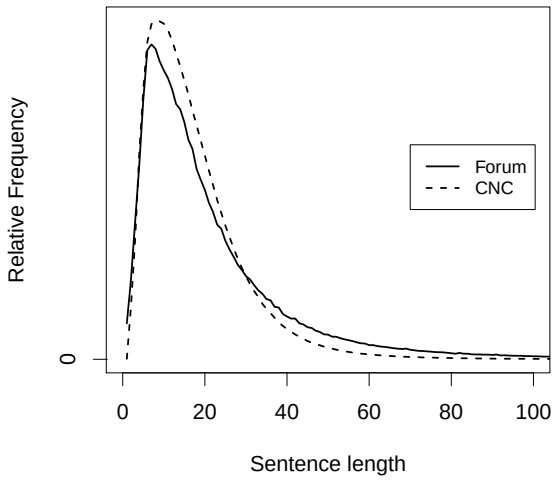

Figure 2: The CZ forum and The Czech National Corpus - sentence length distributions.

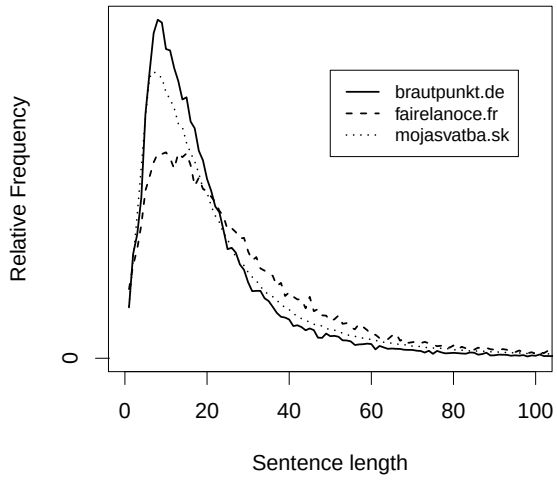

Figure 3: The DE, FR and SK fora - sentence length distributions.

5 Technical issues

Of course, language mutations of the same forum (sharing the same category structure and running on the same forum engine) are a ”researcher’s dream” and not the case of the majority of potential comparable fora.

You will probably ask two questions: 1) How to effectively extract messages from a site with undocumented structure? 2) How to put together comparable fora in multiple languages and how to align their category hierarchy?

5.1 Messages mining

According to an internet source [2], about 96 % of internet discussion fora are powered by two most popular forum systems, phpBB and vBulletin, and another 3 % are powered by Simple Machines Forum, MyBB and Invision Power Board.

Our observation is, that small hobby fora run mostly on unadapted (”as is”) phpBB or another free system, while large commercial fora often have their own systems.

If you intend to automatically process only a few selected fora, you will probably use XPath queries on the HTML Document Object Model. According to our experience, it is very easy and straightforward task to write a single wrapper for a particular forum. But it would be nice, of course, to have a general solution which does not rely on a fixed forum structure. Unfortunately, general web page cleaning algorithms, e.g. Victor (Spousta et al., 2008), are not

[2] http://www.qualityposts.com/ForumMarketShare.php

EN2	How to set up a budget
DE	Budget?
FR2	Financement mariage
CZ	Jaký máte rozpočet na svatbu???
SK	Svadobny rozpocet
EN	Mobile hair and makeup
DE	Friseur und Kosmetik daheim?
FR2	Estheticienne a domicile?
CZ	Nalíčení plus účes doma - Praha
SK	Licenie a uces - v den svadby a doma
EN	Hair extensions?
DE	Echthaar-Clip-Extensions
FR2	Extensions pour cheveux
CZ	Prodlužování vlasů
SK	Predlzovanie vlasov
EN	Where should we go for our honeymoon?
DE	Habt ihr Tipps für eine schöne Hochzeitsreise???
FR2	Quelle destination pour le voyage de noce?
CZ	Svatební cesta
SK	Kam idete na svadobnú cestu?

Table 4: Examples of similar discussions.

very succesfull with this type of input (i.e. ten to fifty rather small textual portions on one page).

However, there are some invariants shared among all types of fora [3]. The content is automatically generated and therefore all the messages on one page (can be generalized to one site) usually "look similar", in terms of HTML structure. (Limanto et al., 2005) exploits this fact and introduces a subtree-matching algorithm for detecting messages on a discussion page. (Li et al., 2009) proposes more complex algorithm which extracts not only the messages content but also the user profile information.

5.2 Fora coupling

The task of optimal fora, categories, discussions, sentences and phrases alignment remains open. Our article is meant to be an inspiration, thus for now, we will not provide our reader with any surprising practical solutions, only with ideas.

The sentence and sub-sentence level can be maintained by existing automatic aligners. For the rest, we believe that combined use of hierarchical structure of the fora together with terms, named entities or simple word translations can help. For example, nearly every EU top level domain hosts a "Volvo Forum" or "Volvo Club", and each Volvo Forum contains some portion of discussions mentioning model names, such as V70 or S60, in their titles.

Besides, according to our case study, the amount of acquired data compared to the amount of human effort should be reasonable even when coupling the fora sites and their top categories manually. Present day approaches to acquiring comparable corpora also require some human knowledge and effort, e.g. you need to pick out manually the most reliable and appropriate news resources.

6 Conclusion

We have proposed an idea of using co-existent web discussion fora in multiple languages addressing the same topic as comparable corpora. Our case study shows that using this approach, one can acquire large portions of comparable multilingual data with minimal effort. We also discussed related technical issues.

You may ask, whether the forum language is the right (addition to a) training set for a machine translation system. The answer may depend on, what type of system it is and what type of input do you want to translate. If you need to translate parliamentary proceedings, you will surely be more satisfied with parliament-only training data. But do you want an anything-to-speech machine translation system to talk to you like a parliamentary speaker, or like a Wikipedia author, or like a friend of yours from your favourite community of interest?

We hope that our article drew the attention of the linguistic audience to this promising source of comparable texts and we are looking forward to seeing some interesting resources and applications.

Acknowledgments

The research described here was supported by the project *GA405/09/0278* of the Grant Agency of the Czech Republic.

[3] and some other types of web sites, eg. e-shops or blogs

References

CNC, 2005. *Czech National Corpus – SYN2005*. Institute of Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic.

Elena Filatova. 2009. Directions for exploiting asymmetries in multilingual wikipedia. In *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies*, CLIAWS3 '09, pages 30–37, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pablo Gamallo and Isaac González López. 2010. Wikipedia as multilingual source of comparable corpora. In *Proceedings of the LREC Workshop on Building and Using Comparable Corpora*, pages 30–37.

Suke Li, Liyong Tang, Jianbin Hu, and Zhong Chen. 2009. Automatic data extraction from web discussion forums. *Frontier of Computer Science and Technology, Japan-China Joint Workshop on*, 0:219–225.

Hanny Yulius Limanto, Nguyen Ngoc Giang, Vo Tan Trung, Jun Zhang, Qi He, and Nguyen Quang Huy. 2005. An information extraction engine for web discussion forums. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, WWW '05, pages 978–979, New York, NY, USA. ACM.

Belinda Maia. 2003. What are comparable corpora? In *Proceedings of the Workshop on Multilingual Corpora: Linguistic requirements and technical perspectives, at the Corpus Linguistics 2003*, pages 27–34, Lancaster, UK, March.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4).

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia, July. Association for Computational Linguistics.

Dragos Stefan Munteanu, Alexander Fraser, and Daniel Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *HLT-NAACL 2004: Main Proceedings*, pages 265–272, Boston, Massachusetts, USA, May. Association for Computational Linguistics.

Miroslav Spousta, Michal Marek, and Pavel Pecina. 2008. Victor: the web-page cleaning tool. In *Proceedings of the Web as Corpus Workshop (WAC-4)*, Marrakech, Morocco.

Jesús Tomás, Jordi Bataller, Francisco Casacuberta, and Jaime Lloret. 2008. Mining wikipedia as a parallel and comparable corpus. *Language Forum*, 34.