Construction and Analysis of a Large Vietnamese Text Corpus

Dieu-Thu Le, Uwe Quasthoff
IMS - University of Stuttgart, NLP Department - University of Leipzig
dieu-thu.le@ims.uni-stuttgart.de, quasthoff@informatik.uni-leipzig.de

Abstract

This paper presents a new Vietnamese text corpus which contains around 4.05 billion words. It is a collection of Wikipedia texts, newspaper articles and random web texts. The paper describes the process of collecting, cleaning and creating the corpus. Processing Vietnamese texts faced several challenges, for example, different from many Latin languages, Vietnamese language does not use blanks for separating words, hence using common tokenizers such as replacing blanks with word boundary does not work. A short review about different approaches of Vietnamese tokenization is presented together with how the corpus has been processed and created. After that, some statistical analysis on this data is reported including the number of syllable, average word length, sentence length and topic analysis. The corpus is integrated into a framework which allows searching and browsing. Using this web interface, users can find out how many times a particular word appears in the corpus, sample sentences where this word occurs, its left and right neighbors.

Keywords: corpus construction, text preprocessing, Vietnamese, topic modeling, searching, word co-occurrences

1. Introduction

Vietnamese text processing started to become active about twelve years ago. Since then, several corpora have been built for some specific natural language processing tasks. (Pham et al., 2007) presented a corpus consisting of newspapers coming from two news sources collected within 6 months in 2005. This corpus was annotated with entity classes such as person, location, organization for named entity recognition. (Tu et al., 2006) released a corpus of 305 newspaper articles together with a list of 2,000 personal names and 707 locations.[1] (Do et al., 2009) prepared a parallel corpus of Vietnamese-French consisting of around 12 million (M) document pairs for machine translation. The EVC corpus produced at the national university of Ho Chi Minh (VNUHCM) consists of 400,000 pairs of English-Vietnamese sentences with approximately 5,500,000 words in the fields of Science and Technology. (Pham et al., 2008) presented a corpus of 1 million words collected from newspapers and children's literature for a comparison in the word uses in children's literature and in general text. SEAlang Library Vietnamese Text Corpus introduced a corpus search interface with word neighbors and sample sentences including more than 79M characters.[2] The Vietnamese corpus project[3] also collected data from newspapers and annotate them with information such as author, public date, register date.

In comparison to other available Vietnamese news corpora, this collection is one of the most comprehensive corpora containing a large amount of text collected from various sources. It can serve as a resource for different Vietnamese natural language processing tasks and text analysis.

[1] http://www.jaist.ac.jp/ hieuxuan/vnwordseg/data/
[2] http://sealang.net/vietnamese/corpus.htm
[3] http://pers-www.wlv.ac.uk/in6930/corpus.htm

2. The construction of the Vietnamese corpus

2.1. Data sources

The Vietnamese corpus is a Web corpus collected between 2007 and 2014. It contains about 70 million of sentences with about 4.05 billion running words. The main sources are Wikipedia (2M sentences), newspaper texts (13M sentences) and randomly crawled web pages (55M). Due to the collection process the actual time of origin for randomly collected texts is impossible to identify. As a rough approximation, the word frequencies for the years 1980-2030 are shown in Figure 1. If we assume that most texts reported online are on the present or recent past, the distribution of these numbers is strongly correlated with the origin of the texts.

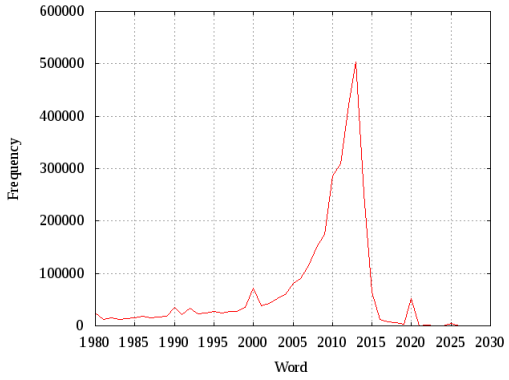

Figure 1: Vietnamese corpus timeline

2.2. Text preprocessing

The corpus pre-processing is described in full detail in (Goldhahn et al., 2012) and results after language identification and text cleaning in sentence separated text without duplicates. After word segmentation and frequency counting word co-occurrences are calculated. All the data are available both for online web searching and for download.

Vietnamese language and problems with word segmentation Tokenization, the process of breaking strings into words, is an important language pre-processing step for further tasks such as parsing, text mining. Since most Vietnamese words are composed of more than one syllable where each syllable is separated by blanks (Dinh et al., 2008), using common tokenizers such as replacing blanks with word boundaries does not work for Vietnamese. As reported in (Hong Phuong et al., 2008), there are about 82% syllables in Vietnamese are words themselves, which correspond to 16% of total Vietnamese words. 71% of words are composed of two syllables, 14% have at least three syllables. There are many syllables that are words themselves, but can also be parts of other words. For example, "bóng" can be *shadow*, while "bóng đèn" means *bulb*; "mặt" is *face* but "mặt trăng" is *moon*.

A review of Vietnamese tokenizers Most studies in this field employ statistical methods such as using probabilistic models (Le Trung et al., 2010), conditional random fields (CRF) and support vector machine (SVM) (Tu et al., 2006). (Tu et al., 2006) considered the problem of detecting word boundaries in a sentence is modeled as that of labeling each syllables as either I_W (inside a word), B_W (begin a word) and O (outside a word). The segmentation tool trained on about 8,000 sentences using CRF and is available online with the name JVnSegmenter[4]. Some other systems use hybrid methods such as in (Pham et al., 2009), (Hong Phuong et al., 2008). (Pham et al., 2009) exploit part-of-speech (POS) information based on maximum matching algorithm combined with stochastic models to tackle the task of word segmentation. In (Hong Phuong et al., 2008), the authors combine finite state automata techniques with the maximal matching strategy and regular expression parsing. This algorithm is implemented in the vnTokenizer tool, a tokenizer for Vietnamese texts.[5]
A comparison of available Vietnamese tokenization tools is reported in (Dinh et al., 2008). In particular, the CRF-based tool of the group (Tu et al., 2006), the PVnSeg tool (unknown source) and the hybrid method of (Hong Phuong et al., 2008) are compared using a test corpus containing 1,264 articles from Politics-Society section of a Vietnamese online newspaper "Tuoi Tre", where words have been manually segmented by linguists. The result shows that both vnTokenizer and JVnSegmenter achieve roughly 94% F-measure. In our preprocessing steps, we will use the JVnSegmenter for preparing the Vietnamese corpus.

2.3. Statistical analysis of the corpus

Corpus statistics has two different aspects: General language statistics (in the case of a large and representative corpus) and special corpus analysis (to compare the values measured for the corpus with representative values). We will give two examples here. For a more general discussion see (Eckart et al., 2012).

Word length for different frequency ranges First we are interested in word length, measured both in characters and number of syllables. Due to the special word structure in Vietnamese, these values are computed as follows:

- Word length in characters is calculated without the possible blanks within a word. In most languages, syllable boundaries are not represented by a character, so the same counting strategy is applied for Vietnamese.
- The number of syllables is trivial to count by counting the blanks within a word plus one. But, the correctness of this value clearly depends on the correctness of the word segmentation.
- For the average syllable length, the average is taken per word, i.e. the syllable length per word is averaged.

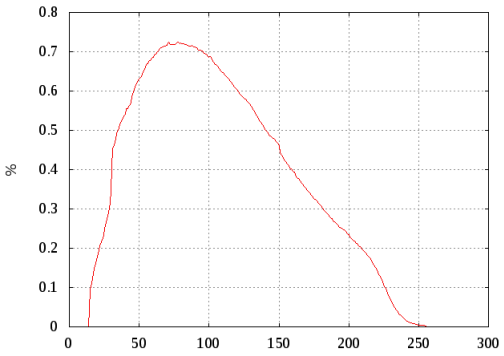

Figure 2: Length of sentences in characters

max (rank)	Word length in characters	Average number of syllable	Average syllable length
1	2.00	1.00	2.00
10	2.90	1.00	2.90
100	3.44	1.07	3.20
1000	4.99	1.45	3.42
10000	5.65	1.63	3.48
100000	6.64	1.75	3.99
1000000	8.04	1.85	4.73

Table 1: Word length in characters (without blank), average number of syllables and average syllable length in characters

The average values for these numbers are calculated for the most frequent N words for N=1, 10, 100, ..., 1.000.000. The average word length increases as expected as a consequence of language economy: Words of higher frequency tend to be shorter. Economy in language also implies an increasing average number of syllables with N. The additional increase in syllable length is due to the fact that for less frequent words some of the syllables tent to become more complex and longer.

Sentence length in characters Figure 2 shows sentence length measured in characters with an average sentence length of 107 characters. While the graph is relatively smooth there are minor irregularities, for instance at length 71 and near length 150. In the case of length 71 the number

[4]http://jvnsegmenter.sourceforge.net/
[5]http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer

of sentences is about 2.000 larger than expected. Manual inspection shows near duplicates of the following form:
Đăng tin Vip: VC VIP 1007974 gửi 8655 (10.000đ) để đăng tin VIP 5 ngày.
Đăng tin Vip: VC VIP 1056511 gửi 8655 (10.000đ) để đăng tin VIP 5 ngày.[6]

These near duplicates, of course, reduce the quality of word frequencies and word co-occurrences. This example illustrate the importance of quality checking during the preprocessing, with a special emphasis on near duplicates.
The frequency data extracted from the corpus together with word related statistics will be used for the Vietnamese Frequency Dictionary to appear in 2016 (Quasthoff et al., 2011).

Topic modeling on the Vietnamese corpus A sample of topics estimated from the Vietnamese corpus using Latent Dirichlet Allocation is illustrated in Table 2. It provides a way of organizing and browsing the data to discover hidden topics within the corpus. Within the same framework, we have also estimated topics for other languages (Figure 3). The topics inferred from different languages could be used to map and compare among languages, for example: to discover widely reported news events around the world using topic timelines, to find out the difference in topics of interests among countries, etc.

2.4. A search interface

We provide a web interface to enable search within the corpus[7]. In figures 4 and 5, the searched word "mai" is ambiguous, it can mean tomorrow, ochna flower, etc. The search results show the number of times this term occurs in the whole corpus, its frequent rank and its frequency class (i.e., the frequency of a word in relation to the most frequent word, the term "mai" appears $\approx 2^7$ times more frequently than the most frequent word of the whole corpus). Some sample sentences where this term occurs are also shown in the search result.
In Figure 5, terms that co-occur most frequently with the given term are shown together with the number of times they co-occur together. Similarly, left and right neighbor cooccurrences of the given term are also listed. An interactive graph shows how the term is associated to its neighbors. From this graph, one can see different meanings and contexts of the searched term (e.g., "ngày mai" refers to tomorrow, "phô mai" is cheese, "mỉa mai" means sarcasm, "mai táng" is burial, "mai sau" expresses later on, etc.).

3. Conclusions

In this paper, we have presented a Vietnamese corpus containing around 4.05 billion words, coming from textual data collected on the internet. We have described main steps for data collection and processing for Vietnamese. From this corpus, we have extracted statistical information such as average word length, number of syllables and syllable length, topic models estimated from the data. A web-interface is also available to search within the corpus, find co-occurrences and examples of sentences where a given word appears.

4. Bibliographical References

Dinh, Q. T., Le, H. P., Nguyen, T. M. H., Nguyen, C. T., Rossignol, M., and Vu, X. L. (2008). Word segmentation of Vietnamese texts: a comparison of approaches. In *6th international conference on Language Resources and Evaluation - LREC 2008*, Marrakech, Morocco. ELRA - European Language Resources Association.

Do, T.-N.-D., Le, V.-B., Bigi, B., Besacier, L., and Castelli, E. (2009). Mining a comparable text corpus for a vietnamese - french statistical machine translation system. In *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*.

Eckart, T., Quasthoff, U., and Goldhahn, D. (2012). The influence of corpus quality on statistical measurements on language resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Hong Phuong, L., Thi Minh Huyen, N., Roussanaly, A., and Vinh, H. T. (2008). A hybrid approach to word segmentation of vietnamese texts. In *Language and Automata Theory and Applications*, volume 5196 of *Lecture Notes in Computer Science*, pages 240–249. Springer Berlin Heidelberg.

Le Trung, H., Le Anh, V., and Le Trung, K. (2010). An unsupervised learning and statistical approach for vietnamese word recognition and segmentation. In *Proceedings of the Second International Conference on Intelligent Information and Database Systems: Part II*, ACIIDS'10, pages 195–204, Berlin, Heidelberg. Springer-Verlag.

Pham, T., Q, T. T., Kawazoe, A., Dinh, D., and Collier, N. (2007). Construction of vietnamese corpora for named entity recognition. In *Conference RIAO*.

Pham, G., Kohnert, K., and Carney, E. (2008). Corpora of vietnamese texts: Lexical effects of intended audience and publication place. In *Behavior Research Methods*.

Pham, D. D., Tran, G. B., and Pham, S. B. (2009). A hybrid approach to vietnamese word segmentation using part of speech tags. In *Proceedings of the 2009 International Conference on Knowledge and Systems Engineering*, KSE '09, pages 154–161, Washington, DC, USA. IEEE Computer Society.

Quasthoff, U., Fiedler, S., and (eds.), E. H. (2011-). Frequency dictionaries. In *Leipziger Universitätsverlag*.

Tu, N. C., Kien, N. T., Hieu, P. X., Minh, N. L., and Thuy, H. Q. (2006). Vietnamese word segmentation with crfs and svms: An investigation. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC 2006)*.

[6] All these sentences come from `http://vico.vn`
[7] http://corpora.informatik.uni-leipzig.de/

khán_giả (audience)	hoa (flower)	tiền (money)	giá (price)
nhạc (music)	vườn (garden)	công_ty (company)	thị_trường (market)
ca_khúc (song)	cây (tree)	nhân_viên (employee)	vàng (gold)
thí_sinh (candidate)	hương (perfume)	hợp_đồng (contract)	nước (country)
âm_nhạc (music)	mùi (smell)	số_tiền (money)	lượng (wage)
ca_sĩ (singer)	giống (species)	lao_động (labor)	Giá (Price)
chương_trình (program)	Đà_Lạt (Dalat)	lương (wage)	thế_giới (world)
đêm (night)	gỗ (wood)	công_việc (job)	mức (level)
ca_sỹ (singer)	mai (ochna tree)	Công_ty (Company)	tăng (increase)
sân_khấu (stage)	sắc (color)	công_nhân (worker)	tuần (week)
tiết_mục (show)	loại (type)	chi (spend)	miếng (piece)
phần (part)	xuân (spring)	khoảng (around)	nhu_cầu (needs)
giám_khảo (examiner)	gốc (root)	sếp (boss)	cung (supply)
show (show)	Cây (Tree)	việc (job)	chuyên_gia (expert)
album (album)	màu (color)	khoản (amount)	đấu_thầu (bidding)
cuộc_thi (competition)	hồng (pink)	thu_nhập (income)	kinh_tế (economic)
nghệ_sĩ (musician)	cánh (petal)	ký (sign)	phiên (auction)
bài_hát (song)	trời (sky)	giám_đốc (director)	thời_điểm (time)
vòng (round)	anh_đào (cherry)	tháng (month)	người_dân (citizen)
màn (performance)	loài (species)	chi_phí (cost)	Nhà_nước (Goverment)

Table 2: Random Vietnamese topics extracted from the corpus

German				English			
Internet	Schüler	Tiere	Haus	game	church	cancer	service
Facebook	Schule	Tier	Wohnung	season	God	disease	network
Daten	Schulen	Natur	Bewohner	team	Church	people	TV
Google	Lehrer	Wald	Wohnungen	games	Francis	health	phone
User	Studenten	Tieren	Fenster	baseball	Catholic	treatment	services
Nutzer	Ausbildung	Menschen	Hauses	home	St.	Dr.	customers
Seite	Klasse	Jagd	Keller	runs	Pope	blood	company
Website	Universität	Pflanzen	Mieter	run	people	study	cable
Netz	Schülern	Baum	Gebäude	inning	Jesus	risk	television
Twitter	Sprache	Arten	Raum	innings	faith	patients	Internet
Werbung	Deutsch	Bäume	Häuser	hits	religion	HIV	phones
Netzwerk	Studium	Umwelt	Tür	League	churches	Health	networks
E-Mail	Schülerinnen	Hektar	Nachbarn	Red	pope	heart	data
Namen	Klassen	Elefanten	Zimmer	manager	world	brain	month
Fotos	Unterricht	Farm	Holz	win	years	cases	access
French				**Spanish**			
santé	film	vendredi	parti	enfants	película	economía	teatro
cas	série	soir	élections	famille	cine	crecimiento	obra
maladie	The	lundi	campagne	femmes	serie	crisis	director
médecins	rôle	jeudi	candidat	femme	director	recuperación	Teatro
traitement	films	dimanche	vote	ans	películas	déficit	espectáculo
cancer	James	samedi	président	père	actor	año	escenario
médecin	réalisateur	mardi	tour	fille	historia	política	música
risque	cinéma	mercredi	voix	mère	personaje	medidas	escena
soins	personnage	matin	Parti	vie	actriz	reformas	ciclo
médicaments	Cinéma	communiqué	candidats	parents	personajes	mercados	público
patients	tournage	journée	pouvoir	fils	papel	países	festival
personnes	Disney	l'AFP	partis	enfant	protagonista	políticas	compañía
Santé	Star	nuit	majorité	hommes	actores	española	danza
risques	scénario	porte-parole	scrutin	couple	cinta	empleo	dirección
maladies	succès	après-midi	mandat	maison	televisión	deuda	montaje
Italian				**Russian**			
vettura	spettacolo	produzione	sera	Украины	спорта	Москве	Михаил
motore	teatro	gas	domenica	Украине	мира	Москвы	Грузии
versione	Teatro	energia	giorni	Украина	России	РИА	Михаила
litri	palco	rifiuti	lunedì	Израиля	место	Новости	Анатолий
cavalli	via	vino	sabato	Киеве	участие	ссылкой	словам
potenza	musica	impianti	giornata	Тимошенко	спортсменов	Интерфакс	Саакашвили
cambio	scena	anni	venerdì	Израиль	соревнования	столицы	время
modello	Festival	tonnellate	pomeriggio	Янукович	соревнований	Сергей	Анатолия
CV	serata	prezzo	giovedì	Виктор	медали	источник	рамках
BMW	concerto	emissioni	mattina	время	чемпионате	представитель	Осетии
km/h	pubblico	territorio	giorno	Украину	чемпионат	столице	обыски
sistema	sabato	raccolta	mercoledì	Киев	турнир	словам	Тбилиси
benzina	tour	benzina	martedì	Израиле	Кубка	Подмосковье	качестве
guida	edizione	fonti	programma	Азаров	спортсмены	агентства	года
velocità	festival	costi	settimana	Киева	соревнованиях	Москва	Грузия

Figure 3: Random topics estimated on other corpora of other languages within the same framework

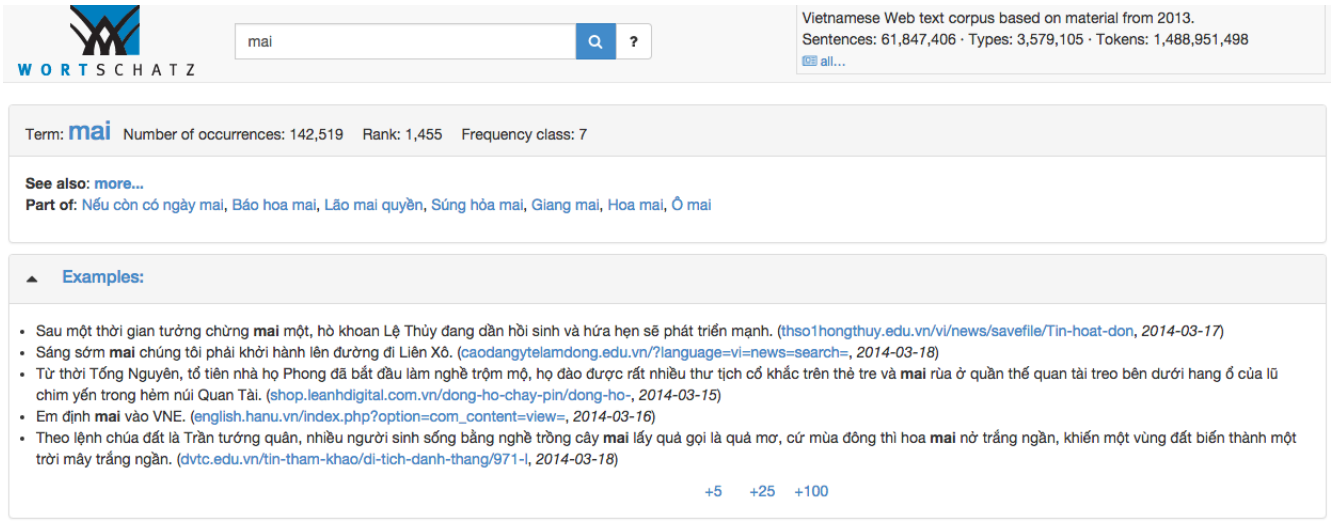

Figure 4: Searching within the Vietnamese corpus: searching for the ambiguous word “mai” (tomorrow, ochna flower)

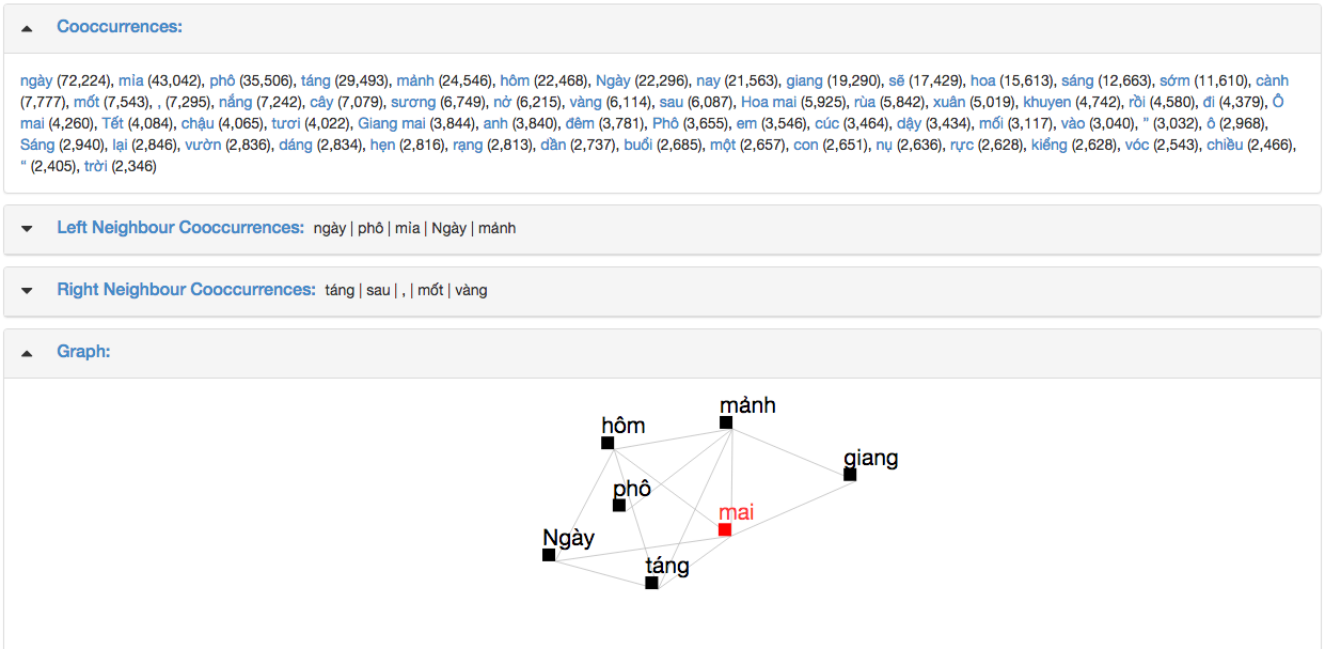

Figure 5: Searching within the Vietnamese corpus: cooccurrences