CIST System Report for ACL MultiLing 2013

-- Track 1: Multilingual Multi-document Summarization

Lei Li, Wei Heng, Jia Yu, Yu Liu, Shuhong Wan
Center for Intelligence Science and Technology (CIST),
School of Computer Science and Technology,
Beijing University of Posts and Telecommunications (BUPT), China
leili@bupt.edu.cn

Abstract

This report provides a description of the methods applied in CIST system participating ACL MultiLing 2013. Summarization is based on sentence extraction. hLDA topic model is adopted for multilingual multi-document modeling. Various features are combined to evaluate and extract candidate summary sentences.

1 Introduction

CIST system has participated Track 1: Multilingual Multi-document Summarization in ACL MultiLing 2013 workshop. It could deal with all ten languages: Arabic, Chinese, Czech, English, French, Greek, Hebrew, Hindi, Romanian and Spanish. It summarizes every topic containing 10 texts and generates a summary in plain text, UTF8 encoding, less than 250 words.

2 System Design

There have been many researches about multi-document summarization, (Wan et al., 2006; He et al., 2008; Flore et al., 2008; Bellemare et al., 2008; Conroy and Schlesinger, 2008; Zheng and Takenobu, 2009; Louis and Nenkova, 2009; Long et al., 2009; Lin and Chen, 2009; Gong et al., 2010; Darling, 2010; Kumar et al., 2010; Genest and Lapalme, 2010; Jin et al., 2010; Kennedy et al., 2010; Zhang et al., 2011), but less about multilingual multi-document summarization (Leuski et al., 2003; Liu et al., 2011; Conroy et al., 2011; Hmida and Favre, 2011; Das and Srihari, 2011; Steinberger et al., 2011; Saggion, 2011; El-Haj et al., 2011).

This system must be applicable for unlimited topics, we couldn't use topic knowledge. Different topic has different language styles, so we use sentence as the processing unit and summarization method based on sentence extraction. It must also be available for different languages, we couldn't use much specific knowledge for all languages except one or two we understand. We refer to a statistical method, hLDA (hierarchical Latent Dirichlet Allocation (LDA)).

LDA has been widely applied. (Arora and Balaraman, 2008; Krestel et al., 2009). Some improvements have been made. (Griffiths et al., 2005; Blei and Lafferty, 2006; Wang and Blei, 2009). One is to relax its assumption that topic number is known and fixed. Teh et al. (2006) provided an elegant solution. Blei et al. (2010) extended it to exploit the hierarchical tree structure of topics, hDLA, which is unsupervised method in which topic number could grow with the data set automatically. There's no relations between topics in LDA (Blei, 2003), but hLDA could organize topics into a hierarchy, in which higher level topics are more abstractive. This could achieve a deeper semantic model similar with human mind and is especially helpful for summarization. Celikyilmaz (2010) provided a multi-document summarization method based on hLDA with competitive results. However, it has the disadvantage of relying on ideal summaries. To avoid this, the innovation of our work is completely dependent on data and hierarchy to extract candidate summary sentences.

Figure 1 and 2 show the framework for ten languages. Since Chinese Hanzi is different from other languages, we treat it with special processing. But the main modules are the same. The kernel one is constructing an hLDA model[1]. It's language independent.

[1] http://www.cs.princeton.edu/~blei/topicmodeling.html

Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pages 39–44,
Sofia, Bulgaria, August 9 2013.

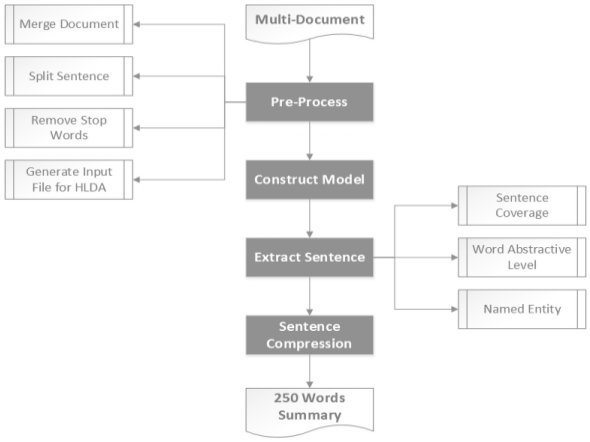

Figure 1: framework for nine languages (no Chinese)

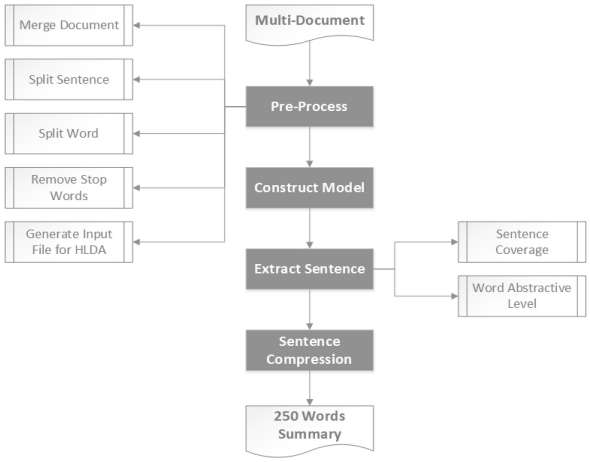

Figure 2: framework for Chinese

3 Text Pre-processing

There are some unified pre-processing steps for all languages and a special step for Chinese.

3.1 Merging Documents

We treat multi-document together, so we firstly combine them into a big text. As to Chinese, we combine and delete empty lines. As to other nine languages, we do this when we split sentences.

3.2 Splitting Sentences

We split sentences to get the processing unit. There are two lines of title and date ending with no punctuation mark. We add a full stop ourselves to avoid them being connected with the first sentence. For Chinese, we split sentences according to ending punctuation marks, while for other nine languages, the full stop "." could have other functions. We adopt machine learning method[2]. After some experiments, we choose Support Vector Machine model for English and French, Naïve Bayes model for other 7 languages.

[2] https://code.google.com/p/splitta/

3.3 Removing Stop Words

We add ICTCLAS[3] word segmentation to Chinese to make all languages have the same word separator. Then we could obtain words easily, among which are some stop words. We construct stop lists. For English and Chinese, the stop list contains punctuation marks and some functional words, while for other languages, it contains punctuation marks, which could unified the whole process easily although generally we do not treat punctuation marks as words. At the same time, all capitalized characters are changed to lower case.

3.4 Generating Input File for hLDA

We build a dictionary for remaining words, which are sorted according to frequency. The more frequent words are located before the less frequent ones. This is a mapping from word to a number varying from 1 to dictionary size. Finally we generate an input file for hLDA, in which each line represents a sentence, in the following form:

[number of words in the sentence] [word-NumberA]:[local frequencyA] [word-NumberB]:[local frequencyB]...

Figure 3 shows an example. As we can see that now it's language independent.

```
5 1:1 6:1 3:1 355:1 205:1
2 188:1 518:1
16 44:1 16:1 172:1 13:1 5:1 1:1 2:1 6:1 207:1 197:1 768:1 794:1 355:1
. 39:1 3:1 231:1
8 6:1 65:1 949:1 227:1 333:1 19:1 20:1 247:1
18 691:1 162:1 41:1 433:1 493:1 0:1 395:1 8:1 11:1 15:1 32:1 4:1 646:1
. 183:1 726:1 541:1 382:1 27:1
```

Figure 3: hLDA input file

4 hLDA Topic Modeling

Given a collection of sentences in the input file, we wish to discover common usage patterns or topics and organize them into a hierarchy. Each node is associated with a topic, which is a distribution across words. A sentence is generated by choosing a path from the root to a leaf, repeatedly sampling topics along that path, and sampling the words from the selected topics. Sentences sharing the same path should be similar to each other because they share the same sub-topics. All sentences share the topic distribution associated with the root node.

As to this system, we set hierarchy depth to 3, because we have found out in former experiments that 2 is too simple, and 4 or bigger is too complex for the unit of sentence.

[3] http://www.nlpir.org/download/ICTCLAS2012-SDK-0101.rar

4.1 Hierarchy Evaluation

In order to make sure that a hierarchy is good, we need to evaluate its performance. The best method is human reading, but it's too laborious to browse all topics and all languages. In fact, we could not understand all ten languages at all. So we build another simpler and faster evaluation method based on numbers. According to former empirical analysis, if a hierarchy has more than 4 paths and the sentence numbers for all paths appear in balanced order from bigger to smaller, and the sentences in bigger paths could occupy 70-85% in all sentences, then we could possibly infer that this hierarchy is good.

4.2 Parameter Setting

When facing a new corpus, we could hardly set the parameters automatically either by human or machine. There is a choice of sampling. We tried it for all languages with 100000 iterations. But the results are poor, even in the worst case each sentence is set to a single path. Thus we give up sampling and try to set the parameters by human.

We begin with Chinese because it seems to be the most difficult case. We randomly choose two topics for original testing and set some parameters according to former experience. Then we evaluate the result using method in 4.1. If it's not good, we go on to adjust the settings until we obtain a satisfactory result. The satisfied settings are then used originally for the whole corpus. Table 1 shows the details.

Parameter	Setting
ETA	1.2 0.5 0.05
GAM	1.0 1.0
GEM_MEAN	0.5
GEM_SCALE	100
SCALING_SHAPE	1.0
SCALING_SCALING	0.5
SAMPLE_ETA	0
SAMPLE_GAM	0

Table 1: Original parameter settings

Language	Topic
English	M006
Hebrew	M001 M006
Romanian	M002
Spanish	M003
Chinese	M004 M006

Table 2: original bad result

After running the whole corpus, we evaluate the results again. We found out that for most cases, the hierarchy is good, but there are some cases not so good, as shown in Table 2. So one set of parameter settings could not deal with all languages and topics successfully. The reason may be that different language and different topic must have different inherent features.

4.3 Parameter Adjustment

We analyze the bad results and try to adjust the settings. For instance, in English M006, there are only two paths indicating that the tree is too clustered. Parameter ETA should be reduced to separate more sub-topics. But too small ETA may lead to hLDA failure without level assignment result in limited iterations. So we also adjust GEM to get closer to the prior explanation of corpus. In some case, the numbers are assigned too much to the former big paths, then we should adjust SCALING parameters to separate some numbers to the smaller paths. For the bad cases in Table 2, we finally use the settings in Table 3.

Parameter	Setting
ETA	5.2 0.005 0.0005
GAM	1.0 1.0
GEM_MEAN	0.35
GEM_SCALE	100
SCALING_SHAPE	2.0
SCALING_SCALING	1.0
SAMPLE_ETA	0
SAMPLE_GAM	0

Table 3: Adjusted parameter settings

Figure 4 shows an example of the modeling result of M004 in English.

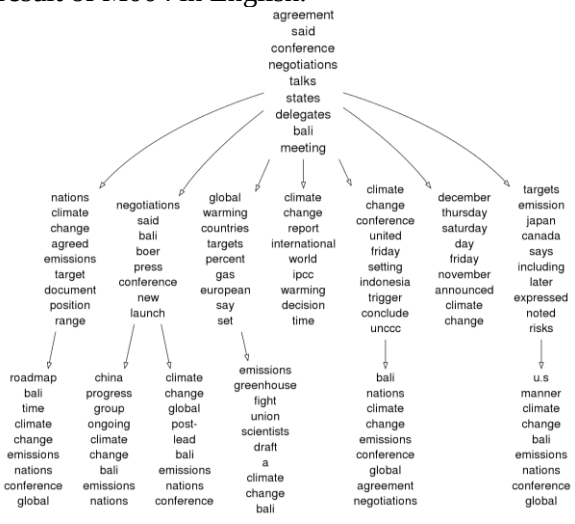

Figure 4: hLDA result example

5 Summary Generation

5.1 Sentence Evaluation

In the hLDA result, sentences are clustered into sub-topics in a hierarchical tree. A sub-topic is more important if it contains more sentences. Trivial sub-topics containing only one or two sentences could be neglected. Final summary

should cover those most important sub-topics with their most representative sentences. We evaluate the sentence importance in a sub-topic considering three features.

1) Sentence coverage, which means that how much a sentence could contain words appearing in more sentences for a sub-topic. We consider sentence coverage of each word in one sentence. The sentence weight is calculated as eq.(1).

$$S_{tf} = \frac{\sum_{i=1}^{|s|} \frac{num_s(w_i)}{n}}{|s|} \quad (1)$$

Where w_i is the i_{th} word in sentence s, $num_s(w_i)$ is the number of sentences that w_i covers, | s | is the number of words in the sentence, and n is the total number of all sentences.

2) Word Abstractive level. hLDA constructs a hierarchy by positioning all sentences on a three-level tree. Level 0 is the most abstractive one, level 2 is the most specific one, and level 1 is between them. We evaluate the sentence abstractive feature as eq.(2).

$$S_l = a * \frac{num(W_0)}{|S|} + b * \frac{num(W_1)}{|S|} + c * \frac{num(W_2)}{|S|} \quad (2)$$

Where $num(W_0)$, $num(W_1)$, $num(W_2)$ are numbers of level 0, 1 and 2 words respectively in the sentence. There are three parameters: a, b and c, which are used to control the weights for words in different levels. Although we hope the summary to be as abstractive as possible, there is really some specific information we also want. For instance, earthquake news needs specific information about death toll and money lost.

3) Named entity. We consider the number of named entities in one sentence. This time we only have time to use Stanford's named entity recognition toolkit[4], which could identify English person, address and institutional names. If one sentence contains more entities, then it has a high priority to be chosen as candidate summary sentence. Let S_n be the number of named entity categories in one sentence. For example, if one sentence has only person names, then S_n is 1; else if it also has address information, then S_n is 2; else if it contains all three categories, then S_n is 3.

At last, we calculate sentence score S as eq. (3, 4), where d, e and f are feature weights:

English: $S = d * S_{tf} + e * S_l + f * S_n$ (3)

Others: $S = d * S_{tf} + e * S_l$ (4)

After experiments, we set {a, b, c, d, e, f} to {0.3, 1, 0.3, 2, 1, 0.05} for English, {a, b, c, d, e} to {1, 0.75, 0.25, 2, 1} for Chinese without M004 and M006, and {0.3, 1, 0.3, 2, 1} for others.

5.2 Summary Generation

We extract 30 candidate sentences with high S ordered by S from bigger to smaller and check them one by one. We use 30 sentences to make sure that when a candidate sentence is not good to be in a final summary, we could have enough other alternative sentences with less S. Then we generate the final summary as Figure 5.

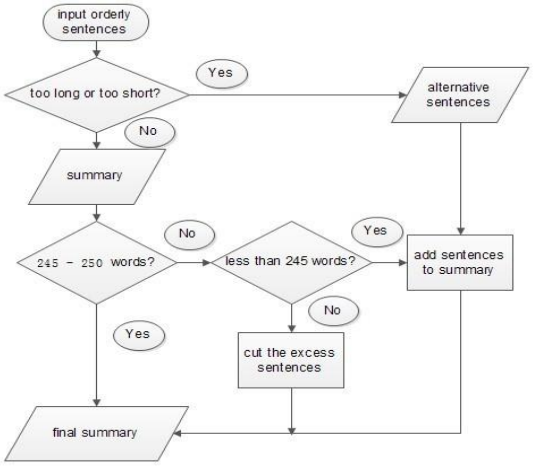

Figure 5: 250-summary generation flow chart

6 Evaluations

We've got only the automatic evaluation result. CIST could get best performance in some language, such as Hindi in ROUGE, and in some topics, such as Arabic M104, English and Romania M005, Czech M007, Spanish M103 etc. in N-gram graph methods: AutoSummENG, MeMoG and NPowER. CIST could also get nearly worst performance in some cases, such as French and Hebrew. In other cases it gets middle performance. But Chinese result looks very strange to us; we think that it needs more special discussion.

7 Conclusion and Future Work

hLDA is a language independent model. It could work well sometimes, but not stable enough. Future work will focus on parameter adjustment, modeling result evaluation, sentence evaluation and good summary generation.

Acknowledgments

We get support from NSFC 61202247, 71231002, Fundamental Research Funds for Central Universities 2013RC0304 and Beijing Science and Technology Information Institute.

[4] http://nlp.stanford.edu/software/CRF-NER.shtml

References

Abdullah Bawakid and Mourad Oussalah, 2008. *A Semantic Summarization System: University of Birmingham at TAC 2008. TAC 2008 Proceedings*.

Alistair Kennedy, Terry Copeck, Diana Inkpen and Stan Szpakowicz, 2010.*Entropy-based Sentence Selection with Roget's Thesaurus. TAC 2010 Proceedings.*

Annie Louis and Ani Nenkova, 2009. *Predicting Summary Quality using Limited Human Input. TAC 2009 Proceedings.*

Anton Leuski, Chin-Yewlin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard Hovy,2003. *Cross-Lingual C*ST*RD: English Access to Hindi Information. ACM Transactions on Asian Language Information Processing,* 2(3):245–269.

Arora Rachit, and Balaraman Ravindran, 2008. *Latent dirichlet allocation based multi-document summarization. Proceedings of the second workshop on Analytics for noisy unstructured text data. ACM, 2008.*

Asli Celikyilmaz and Dilek Hakkani-Tur. 2010. *A hybrid hierarchical model for multi-document summarization. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* 815–824, Uppsala, Sweden, 11-16 July 2010.

Blei D. and Lafferty J., 2006. *Dynamic topic models. In International Conference on Machine Learning (2006).* ACM, New York, NY, USA:113–120.

Blei D., Griffiths T. and Jordan M., 2010. *The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. J. ACM* 57, 2 (2010):1–30.

Chin-Yew Lin and Eduard Hovy, 2002. *Automated Multi-document Summarization in NeATS. Proceedings of HLT 2002, Second International Conference on Human Language Technology Research.*

Chong Long, Minlie Huang and Xiaoyan Zhu, 2009. *Tsinghua University at TAC 2009: Summarizing Multi-documents by Information Distance. TAC 2009 Proceedings.*

D. M. Blei, A. Ng, and M. Jordan. 2003. *Latent dirichlet allocation, Jrnl. Machine Learning Research*, 3:993-1022, 2003b.

Feng Jin, Minlie Huang and Xiaoyan Zhu, 2010.*The THU Summarization Systems at TAC 2010. TAC 2010 Proceedings.*

Firas Hmida and Benoit Favre, 2011. *LIF at TAC Multiling: Towards a Truly Language Independent Summarizer. TAC 2011 Proceedings.*

Griffiths T., Steyvers M., Blei D. and Tenenbaum J., 2005. *Integrating topics and syntax. Advances in Neural Information Processing Systems 17*. L. K. Saul, Y. Weiss, and L. Bottou, eds. MIT Press, Cambridge, MA,2005:537–544.

Hongyan Liu, Ping'an Liu, Wei Heng and Lei Li, 2011.*The CIST Summarization System at TAC 2011. TAC 2011 Proceedings.*

Horacio Saggion, 2011. *Using SUMMA for Language Independent Summarization at TAC 2011. TAC 2011 Proceedings.*

John M. Conroy and Judith D. Schlesinger, 2008. *CLASSY and TAC 2008 Metrics. TAC 2008 Proceedings*.

John M. Conroy, Judith D. Schlesinger and Dianne P. O'Leary, 2006. *Topic-Focused Multi-document Summarization Using an Approximate Oracle Score. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*: 152–159.

John M. Conroy, Judith D. Schlesinger and Jeff Kubina, 2011.*CLASSY 2011 at TAC: Guided and Multi-lingual Summaries and Evaluation Metrics. TAC 2011 Proceedings.*

Jorge Garc á Flores, Laurent Gillard and Olivier Ferret, 2008. *Bag-of-senses versus bag-of-words: comparing semantic and lexical approaches on sentence extraction. TAC 2008 Proceedings*.

Josef Steinberger, Mijail Kabadjov, Ralf Steinberger, Hristo Tanev, Marco Turchi and Vanni Zavarella,2011. *JRC's Participation at TAC 2011: Guided and Multilingual Summarization Tasks. TAC 2011 Proceedings.*

Judith D. Schlesinger, Dianne P. O'Leary and John M. Conroy, 2008. *Arabic/English Multi-document Summarization with CLASSY—The Past and the Future. CICLing 2008 Proceedings*: 568–581.

Krestel Ralf, Peter Fankhauser and Wolfgang Nejdl, 2009. *Latent dirichlet allocation for tag recommendation. Proceedings of the third ACM*

conference on Recommender systems. ACM, 2009.

Mahmoud El-Haj, Udo Kruschwitz and Chris Fox, 2011. *University of Essex at the TAC 2011 Multilingual Summarisation Pilot. TAC 2011 Proceedings.*

Niraj Kumar, Kannan Srinathan and Vasudeva Varma, 2010. *An Effective Approach for AESOP and Guided Summarization Task. TAC 2010 Proceedings.*

Pierre-Etienne Genest and Guy Lapalme, 2010.*Text Generation for Abstractive Summarization. TAC 2010 Proceedings.*

Pradipto Das and Rohini Srihari, 2011.*Global and Local Models for Multi-Document Summarization. TAC 2011 Proceedings.*

Renxian Zhang, You Ouyang and Wenjie Li, 2011.*Guided Summarization with Aspect Recognition. TAC 2011 Proceedings.*

Shih-Hsiang Lin and Berlin Chen, 2009. *THE NTNU SUMMARIZATION SYSTEM AT TAC 2009. TAC 2009 Proceedings.*

Shu Gong, Youli Qu and Shengfeng Tian, 2009. *Summarization using Wikipedia. TAC 2010 Proceedings.*

Sylvain Bellemare, Sabine Bergler and Ren ́e Witte, 2008. *ERSS at TAC 2008. TAC 2008 Proceedings.*

Teh Y., Jordan M., Beal M. and Blei D., 2006. *Hierarchical Dirichlet processes. J. Am. Stat. Assoc.* 101, 476(2006):1566–1581.

Tingting He, Jinguang Chen, Zhuoming Gui, and Fang Li, 2008. *CCNU at TAC 2008 : Proceeding on Using Semantic Method for Automated Summarization Yield. TAC 2008 Proceedings.*

Wang C. and Blei D., 2009. *Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. Advances in Neural Information Processing Systems 22.* Y. Bengio, D. Schuurmans, J. Lafferty, C.

William M. Darling, 2010. *Multi-Document Summarization from First Principles. TAC 2010 Proceedings.*

Xiaojun Wan, Jianwu Yang and Jianguo Xiao, 2006. *Using Cross-Document Random Walks for Topic-Focused Multi-Document Summarization. WI 2006 Main Conference Proceedings.*

Yuanrong Zheng and Tokunaga Takenobu, 2009. *The TITech Summarization System at TAC-2009. TAC 2009 Proceedings.*