

Abstractive Multi-document Summarization with Semantic Information Extraction

Wei Li

Key Lab of Intelligent Info. Processing,
Institute of Computing Technology, CAS
Beijing, 100190, China
weili@kg.ict.ac.cn

Abstract

This paper proposes a novel approach to generate abstractive summary for multiple documents by extracting semantic information from texts. The concept of *Basic Semantic Unit (BSU)* is defined to describe the semantics of an event or action. A semantic link network on *BSUs* is constructed to capture the semantic information of texts. Summary structure is planned with sentences generated based on the semantic link network. Experiments demonstrate that the approach is effective in generating informative, coherent and compact summary.

1 Introduction

Most automatic summarization approaches are extractive which leverage only literal or syntactic information in documents. Sentences are extracted from the original documents directly by ranking or scoring and only little post-editing is made (Yih et al., 2007; Wan et al., 2007; Wang et al., 2008; Wan and Xiao, 2009). Pure extraction has intrinsic limits compared to abstraction (Carenini and Cheung, 2008).

Abstractive summarization requires semantic analysis and abstract representation of texts, which need knowledge on and beyond the texts (Zhuge, 2015a). There are some abstractive approaches in recent years: sentence compression (Knight and Marcu, 2000; Knight and Marcu, 2002; Cohn and Lapata, 2009), sentence fusion (Barzilay and McKeown, 2005; Filippova and Strube, 2008), and sentence revision (Tanaka et al., 2009). However, these approaches are sentence rewriting techniques based on syntactical

analysis without semantic analysis and abstract representation.

Fully abstractive summarization approach requires a separate process for the analysis of texts that serves as an intermediate step before the generation of sentences (Genest and Lapalme, 2011). Statistics of words or phrases and syntactical analysis that have been widely used in existing summarization approaches are all shallow processing of text. It is necessary to explore summarization methods based on deeper semantic analysis.

We define the concept of *Basic Semantic Unit (BSU)* to express the semantics of texts. A *BSU* is an action indicator with its obligatory arguments which contain actor and receiver of the action. *BSU* is the most basic element of coherent information in texts, which can describe the semantics of an event or action. The semantic information of texts is represented by extracting *BSUs* and constructing *BSU* semantic link network (Zhuge, 2009). Semantic Link Network consists of semantic nodes, semantic links and reasoning rules (Zhuge, 2010; 2011; 2012; 2015b). The semantic nodes can be any resources. In this work, the semantic nodes are *BSUs* extracted from texts. We use semantic relatedness between *BSUs* as semantic links. Then summary can be generated based on the semantic link network through summary structure planning.

The characteristics of our approaches are as follows:

- Each *BSU* describes the semantics of an event or action. The semantic relatedness between *BSUs* can capture the context semantic relations of texts.
- The *BSU* semantic link network is an abstract representation of texts. Reduction on the network can obtain important information of texts with no redundancy.

- Summary is built from sentence to sentence to a coherent body of information based on the *BSU* semantic link network by summary structure planning.

2 Related Work

There are some abstractive summarization approaches in recent years. An approach *TTG* attempts to generate abstractive summary by using text-to-text generation to generate sentence for each subject-verb-object triple (Genest and Lapalme, 2011). A system that attempts to generate abstractive summaries for spoken meetings was proposed (Wang and Cardie, 2013). It identifies relation instances that are represented by a lexical indicator with an argument constituent from texts. Then the relation instances are filled into templates which are extracted by applying multiple sequence alignment. Both of these systems need to select a subset of the large volumes of generated sentences. However, our system generates summary directly by summary structure planning. It can generate well-organized and coherent summary more effectively.

A recent work aims to generate abstractive summary based on Abstract Meaning Representation (AMR) (Liu et al., 2015). It first parses the source text into AMR graphs, and then transforms them into a summary graph and plans to generate text from it. This work only focuses on the graph-to-graph transformation. The module of text generation from AMR has not been developed. The nodes and edges of AMR graph are entities and relations between entities respectively, which are sufficiently different from the *BSUs* semantic link network. Moreover, texts can be generated efficiently from the *BSUs* network. Another recent abstractive summarization method generates new sentences by selecting and merging phrases from the input documents (Bing et al., 2015). It first extracts noun phrases and verb-object phrases from the input documents, and then calculates saliency scores for them. An ILP optimization framework is used to simultaneously select and merge informative phrases to maximize the salience of phrases and meanwhile satisfy the sentence construction constraints. As the results show that the method is difficult to generate new informative sentences really different from the original sentences and may generate some none factual sentences since phrases from different sentences are merged.

Open information extraction has been proposed by (Banko et al., 2007; Etzioni et al.,

2011). They extract binary relations from the web, which is different from our approach that extracts events or actions expressed in texts.

3 The Summarization Framework

Our system produces an abstractive summary for a set of topic related documents. It consists of two major components: Information extraction and summary generation.

3.1 Information Extraction

The semantic information of texts is obtained by extracting *BSUs* and constructing *BSU* semantic link network. A *BSU* is represented as an **actor-action-receiver** triple, which can both detects the crucial content and incorporates enough syntactic information to facilitate the downstream sentence generation. Some actions may not have the receiver argument. For example, “*Flight MH370 – disappear*” and “*Flight MH370 - leave - Kuala Lumpur*” are two *BSUs*.

***BSU* Extraction.** *BSUs* are extracted from the sentences of the documents. The texts are pre-processed by name entity recognition (Finkel et al., 2005) and co-reference resolution (Lee et al., 2011). Constituent and dependency parses are obtained by Stanford parser (Klein and Manning, 2003). The eligible action indicator is restricted to be a predicate verb; the eligible actor and receiver arguments are noun phrase. Both the actor and receiver arguments take the form of constituents in the parse tree. A valid *BSU* should have one action indicator and at least one actor argument, and satisfy the following constraints:

- The actor argument is the nominal subject or external subject or the complement of a passive verb which is introduced by the preposition “by” and does the action.
- The receiver argument is the direct object or the passive nominal subject or the object of a preposition following the action verb.

We create some manual rules and syntactic constraints to identify all *BSUs* based on the syntactic structure of sentences in the input texts.

Constructing *BSU* Semantic Link Network.

The semantic relatedness between *BSUs* contains three parts: *Arguments Semantic Relatedness* (ASR), *Action-Verbs Semantic Relatedness* (VSR) and *Co-occurrence in the Same Sentence* (CSS). Arguments of *BSUs* include actors and receivers, which both are noun phrases and indicate concepts or entities in the text. When computing ASR, the semantic relatedness between concepts must be measured. We use the explicit

semantic analysis based on Wikipedia to compute semantic relatedness between concepts (Gabrilovich and Markovitch, 2007). When computing VSR, WordNet-based measure is used to calculate the semantic relatedness between action verbs (Mihalcea et al., 2006). CSS is measured whether two different *BSUs* co-occur in the same sentence. Semantic relations between *BSUs* are computed by linearly combining these three parts. Then *BSUs* that are extracted from the texts form a semantic link network.

Semantic Link Network Reduction. A discriminative ranker based on Support Vector Regression (SVR) (Smola and Scholkopf, 2004) is utilized to assign each *BSU* a summary-worthy score. Training data was constructed from the DUC 2005 datasets which contain both the source documents and human generated reference summaries. *BSUs* are extracted from these datasets. For each *BSU* in the source documents, if it has occurred in the corresponding human generated summaries or the semantic relatedness between the *BSU* and one *BSU* in the corresponding human generated summaries is above a threshold δ , then it is considered to be a positive sample and be assigned 1 to its summary-worthy score. Otherwise, the *BSU* is considered to be a negative sample and be assigned 0 to its summary-worthy score. Table 1 displays the features of *BSU* used in the SVR model. Then the saliency score of each *BSU* in the semantic link network is calculated by the following equation:

$$Sal(BSU_i) = SW_i * \sum_j R_{ij} \quad (1)$$

Where SW_i is the summary-worthy score of BSU_i ; R_{ij} is the semantic relatedness between BSU_i and BSU_j .

BSUs in the semantic link network are clustered by hierarchical complete-link clustering methods. *BSUs* in each cluster are semantically similar. For example, *Malaysia Airlines plane - vanish* and *Flight MH370 - disappear*. Only the most important one with the largest saliency score is reserved in the network. These less important *BSUs* are eliminated. The remaining *BSU* semantic link network represents the important information of the texts with no redundancy.

3.2 Summary Generation

The summary for the documents is generated directly based on the *BSU* semantic link network. The summary should be well-structured and well-organized. It should not just be a heap of related information, but should build from sen-

Basic Features
Number of words in actor/receiver
Number of nouns in actor/receiver
Number of new nouns in actor/receiver
Actor/receiver has capitalized word?
Actor/receiver has stopword?
Action is a phrasal verb?
Content Features
Actor/receiver has name entity?
TF/IDF/TF-IDF of action
TF/IDF/TF-IDF min max average of actor/receiver
Syntax Features
Constituent tag of actor/action/receiver
Dependency relation of action with actor
Dependency relation of action with receiver

Table 1. Features for *BSU* summary-worthy scoring. We use SVM-light with RBF kernel by default parameters (Joachims, 1999).

tence to sentence to a coherent body of information about a topic.

The summary structure is planned based on the *BSU* semantic link network. An optimal path which covers all the nodes in the network is found. The following two factors are considered when finding the optimal path: (1) **Context Semantic Coherent.** To make the summary semantic coherent, all adjacent sentences should be semantically related. We need to find an optimal path, in which every two adjacent nodes are strong semantically related. The optimal path is denoted as $P=[p_1, p_2, ..., p_n]$ and maximize

$1/n \sum_{i=1}^{n-1} R_{p_i p_{i+1}}$. (2) **Clear-cut Theme.** To make the theme of generated summary clear-cut, the important content should be put in prior position. The order of the i th node in the path is denoted as u_i and its weight is denoted as $w_i = 1 / Sal(BSU_i)$ and maximize $\sum_{i=1}^n u_i w_i$.

To combine the above two factors, we need to find an optimal path which covers each node only once and has the longest distance. The biased-sum weight of all nodes in the path should be maximized. The problem can be proved to be NP-hard by reduction to TSP problem. It can be formalized as an integer linear programming (ILP) as follow. x_{ij} is defined to indicate whether the optimal path goes from node i to node j .

$$x_{ij} = \begin{cases} 1 & \text{if the path goes from node } i \text{ to node } j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Since each node can be traversed only once, the following constraints must be satisfied.

$$\begin{aligned} \sum_{i=1, j \neq i}^n x_{ij} &= 1 & j &= 1, \dots, n \\ \sum_{j=1, j \neq i}^n x_{ij} &= 1 & i &= 1, \dots, n \end{aligned} \quad (3)$$

The nodes in the path are sequentially ordered. If the edge between two nodes is in the path, then

System	ROUGE-1	ROUGE-2	ROUGE-SU4
<i>OurSystem</i>	0.42145	0.11016	0.15632
<i>MultiMR</i>	0.41967	0.10302	0.15385
<i>RankBSU</i>	0.39123	0.08742	0.14381
<i>TTG</i>	0.39268	0.09645	0.14553
<i>AveDUC</i>	0.39684	0.09495	0.14671
<i>NIST Baseline</i>	0.33126	0.06425	0.11114

Table 2. Comparison results (F-measure) on DUC 2007 under ROUGE evaluation.

System	OurSystem	MultiMR	RankBSU	TTG
<i>Pyr (Th:0.6)</i>	0.858	0.845	0.832	0.834
<i>Pyr (Th:0.65)</i>	0.743	0.731	0.718	0.721

Table 3. Comparison results on DUC 2007 under the automated pyramid evaluation with two threshold value 0.6 and 0.65.

the order of the two nodes is sequentially close to each other, which can be formulated as follow:

$$\begin{aligned}
u_i - u_j + nx_{ij} &\leq n-1 \quad 1 \leq i \neq j \leq n \\
1 \leq u_i &\leq n \quad i = 1, \dots, n \\
u_i &\in \mathbb{Z} \quad i = 1, \dots, n
\end{aligned} \quad (4)$$

At last, we can formulate the objective function as follow:

$$\max 1/n \sum_{i=1}^n \sum_{j=1, j \neq i}^n R_{ij} x_{ij} + \lambda \sum_{i=1}^n w_i u_i \quad (5)$$

where parameter λ tunes the effect of the two parts and n is the quantity of *BSUs* in the final *BSU* semantic link network (after reduction).

Sentence Generation. After the summary structure has been planned, sentences are generated for each node in the *BSU* semantic link network. As the *BSU* contains enough semantic and syntactic information, sentence can be generated efficiently according to the following rules:

- Generate a Noun Phrase (NP) based on the actor argument to represent the subject, a NP based on the receiver argument to represent the object if present.
- Generate a Verb Phrase (VP) based on the action verb to link the components above. The tense of the verb is set to the same as in the original sentence, and most modifiers like auxiliaries and negation are conserved.
- Generate complements for the VP when the *BSU* has no receiver. The verb modifiers following the action verb such as prepositional phrases and infinitive phrases can be used as the complement, in case that the verb would have no interesting meaning without a complement.

The process of sentence generation for each node is based on the syntactic structure of the source sentence where the *BSU* is extracted from. The time and location preposition phrases which

are important information of new events are kept. The generated sentences are organized according to the summary structure. If some adjacent sentences in the summary have the same subject, the subject of the latter can be substituted by a pronoun (such as it or they) to avoid repetition of noun phrases. One sample summary generated by our system for “Malaysia MH370 Disappear” news is shown in Figure 1.

4 Evaluation Results

4.1 Dataset and Experimental Settings

In order to evaluate the performance of our system, we use two datasets that have been used in recent multi-document summarization shared tasks: DUC2005 and DUC2007. Each task has a gold standard dataset consisting of document clusters and reference summaries. In our experiments, DUC2005 was used for training and parameter tuning, and DUC2007 was used for testing. Based on the tuning set, the parameter λ is set as 10 and δ is set as 0.7 after tuning.

Our system is compared with one state-of-the-art graph-based extractive approach *MultiMR* (Wan and Xiao, 2009) and one abstractive approach *TTG* (Genest and Lapalme, 2011). In addition, we have implemented another baseline *RankBSU* which uses the graph-based ranking methods on the *BSUs* network to rank *BSUs* and select the top ranked *BSUs* to generate sentences.

4.2 Results

ROUGE-1.5.5 toolkit was used to evaluate the quality of summary on DUC 2007 dataset (Lin and Hovy, 2003). The ROUGE scores of the NIST Baseline system (i.e. *NIST Baseline*) and average ROUGE scores of all the participating systems (i.e. *AveDUC*) for DUC 2007 main task were also listed. According to the results in Table 2, our system much outperforms the *NIST Baseline* and *AveDUC*, and achieves higher ROUGE scores than the abstractive approach *TTG*. So the abstract representation of texts and the information extraction process in our system are effective for multi-document summarization. Our system also achieves better performance than the baseline *RankBSU*, which demonstrates that the network reduction method is more efficient than the popular graph-based ranking methods. As compared with the state-of-art graph-based extractive method *MultiMR*, our system also achieves better performance. Furthermore, our system is abstractive with abstract representation and sentence generation. Incorrect

parser and co-reference resolution will lead to wrong extraction of *BSU*. If with more accurate parser and co-reference resolution, our system will be expected to achieve better performance.

Since ROUGE metric evaluates summaries only from word overlapping perspective, we also use the pyramid evaluation metric (Nenkova and Passonneau, 2004) which can measure the summary quality beyond simply string matching. The pyramid evaluation metric involves semantic matching of summary content units (SCUs) so as to recognize alternate realizations of the same meaning, which is a better metric for the abstractive summary evaluation. Since the manual pyramid evaluation is time-consuming and the evaluation results can't be reproducible with different groups of assessors, we use the automated version of pyramid proposed in (Passonneau et al., 2013) and adopt the same setting as in (Bing et al., 2015). Table 3 shows the evaluation results of our system and the three baseline systems on DUC 2007. The results show that the performance of our system is significantly better than the three baseline systems, which demonstrates that the summaries of our system contain more SCUs than summaries of other systems. So our system can generate more informative summary.

In addition, large volumes of news texts for popular news events are crawled from the news websites. Figure 1 and 2 show the summaries for the “Malaysia MH370 Disappear” news event generated by our system and *MultiMR* respectively. The summary by *MultiMR* contains some repetition of facts obviously. And it is just a heap of information about MH370. The summary by our system doesn't contain much repetition of facts, so it can contain more useful information. And it is built from sentence to sentence to a coherent body. Obviously, the summary by our system is more coherent and compact.

5 Conclusions and Future Works

The proposed summarization approach is effective in information extraction and achieves good performance on DUC datasets. Through the sample summary, we can find that the approach is very effective for summarizing texts that mainly describe facts and actions of news event. Summaries generated by our system are informative, coherent and compact.

But for texts expressing opinions, the approach can't settle it appropriately. For example, when the verbs of *BSUs* are not meaningful actions, like “be”, the semantic relations between

Flight MH370 disappeared after leaving Kuala Lumpur. It had been expected to land in Beijing at 06:30. It took off at 00:41 MYT from runway 32R. It ended in the southern Indian Ocean. The aircraft, a Boeing 777-200ER made a sharp turn westwards. It passed into Vietnamese airspace. The captain of another aircraft attempted to reach the crew of Malaysia Airlines flight MH370. Malaysia Airlines flight 386 was requested to attempt to contact Malaysia Airlines flight MH370 on the Lumpur Radar frequency. Malaysia Airlines flight MH17, another Boeing 777-200ER, was surpassed Malaysia Airlines flight MH370. Malaysia Airlines assumes beyond reasonable doubt there are no survivors. They reported the Malaysia Airlines flight MH370 missing. They releases passenger manifest of flight MH370. They will give US\$ 5000 to the relatives of each passenger. Malaysia released preliminary report. It set up a Joint Investigation Team. Southeast Asian states have joined forces to search waters between Malaysia and Vietnam. Chinese government criticizes Malaysia for inadequate answers regarding Malaysia Airlines flight MH370. Malaysia will be deploying more ships and equipment to assist in the search. It ends hunt in South China Sea. Continued refinement of analysis of flight MH370's satellite communications identified a wide area search. Australia and Malaysia are working on a Memorandum of Understanding to cover financial and co-operation arrangements for search and recovery activities.

Figure 1. The summary of “Malaysia MH370 Disappear” news event generated by our system.

Malaysia Airlines said in a statement that **flight MH370 had disappeared** at 02:40 local time on Saturday after leaving Kuala Lumpur. Southeast Asian states have joined forces to search waters between Malaysia and Vietnam after a **Malaysia Airlines plane vanished on a flight to Beijing**, with 239 people on board. Flight MH370 had been expected to **land in Beijing** at 06:30. If Malaysia Airlines flight MH370 had impacted the ocean hard, resulting underwater sounds could have been detected by hydrophones, given favorable circumstances. Scientists from the CTBTO analyzed their recordings soon after **flight MH370 disappeared**, finding nothing of interest. The CMST researchers believe that the most likely explanation of the hydroacoustic data is that they come from the same event, but unrelated to Malaysia Airlines flight MH370. The lead researcher of the CMST team, Dr.Alec Duncan, believes there's a slim chance that the acoustic event is related to Malaysia Airlines flight MH370. Several IMOS recorders deployed in the Indian Ocean off northwestern Australia by CMST may have recorded data related to Malaysia Airlines flight MH370. Malaysia Airlines released the names and nationalities of the 227 passengers and 12 crew members, based on the flight manifest, later modified to include two Iranian passengers travelling on stolen passports. If the data relates to the same event, related to flight MH370, but the arc derived from analysis of the aircraft's satellite transmission is incorrect, then the most likely place to look for the aircraft would be along a line from HA01.

Figure 2. The summary of “Malaysia MH370 Disappear” news event generated by *MultiMR*.

them can't be appropriately computed by the methods described in the paper. More efficient methods to computer semantic relations between *BSUs* should be developed in the following work.

The sentence generation process described in the paper is just a preliminary scheme. It should be developed to generate sentence relying less on the original sentence structure and aggregating information from several different *BSUs*.

Reference

- Barzilay, R., and McKeown, K. R. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3): 297-328.
- Banko, M., Cafarella, M. J., Soderland, S., et al. 2007. Open information extraction for the web. In *IJCAI 2007*, 7: 2670-2676.
- Bing, L., Li, P., Liao, Y., Lam, W., et al., 2015. Abstractive Multi-Document Summarization via Phrase Selection and Merging. In *ACL 2015*, 1587-1597

- Carenini, G., and Cheung, J. C. K. 2008. Extractive vs. NLG-based abstractive summarization of evaluative text: The effect of corpus controversiality. *Proceedings of the Fifth International Natural Language Generation Conference*. Association for Computational Linguistics, 33-41.
- Cohn, T., and Lapata, M. 2009. Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*, 637-674.
- Etzioni, O., Fader, A., Christensen, J., et al. 2011. Open Information Extraction: The Second Generation. In *IJCAI 2011*, 11: 3-10.
- Finkel, J. R., Grenager, T., and Manning, C. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *ACL 2005*, 363-370.
- Filippova, K., and Strube, M. Sentence fusion via dependency graph compression. In *EMNLP 2008*, 177-185.
- Gabrilovich, E., and Markovitch, S. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *IJCAI 2007*, 7: 1606-1611.
- Genest, P. E., and Lapalme, G. 2011. Framework for abstractive summarization using text-to-text generation. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, 64-73.
- Joachims, T. 1999. SvmLight: Support vector machine. SVM-Light Support Vector Machine <http://svmlight.joachims.org/>, *University of Dortmund*, 19(4).
- Knight, K., and Marcu, D. 2000. Statistics-based summarization-step one: Sentence compression. In *AAAI/IAAI 2000*, 703-710.
- Knight, K., and Marcu, D. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. In *Artificial Intelligence*. 139(1): 91-107
- Klein, D., and Manning, C. D. 2003. Accurate unlexicalized parsing. In *ACL 2003*, 423-430.
- Liu, F., Flanagan, J., et al. 2015. Toward Abstractive Summarization Using Semantic Representations. In *HLT-NAACL 2015*.
- Lin, C. Y., and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *HLT-NAACL 2003*, 71-78.
- Lee, H., Peirsman, Y., Chang, A., et al. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *ACL 2011*, 28-34.
- Mihalcea, R., Corley, C., and Strapparava, C. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI 2006*, 6: 775-780.
- Nenkova, A., and Passonneau, R. 2004. Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL*, pages 145-152.
- Passonneau, R. J., Chen, E., Guo, W., and Perin, D. 2013. Automated Pyramid Scoring of Summaries using Distributional Semantics. In *ACL(2)*, pages: 143-147.
- Smola, A. J., and Schölkopf, B. 2004. A tutorial on support vector regression. *Statistics and computing*, 14(3): 199-222.
- Tanaka, H., Kinoshita, A., Kobayakawa, T., Kumano, T., and Kato, N. 2009. Syntax-driven sentence revision for broadcast news summarization. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, 39-47.
- Wan, X., Yang, J., and Xiao, J. 2007. Manifold-Ranking Based Topic-Focused Multi-Document Summarization. In *IJCAI 2007*, 7:2903-2908.
- Wang, D., Li, T., Zhu, S., and Chris, D. 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *SIGIR 2008*, 307-314.
- Wan, X., and Xiao, J. 2009. Graph-Based Multi-Modality Learning for Topic-Focused Multi-Document Summarization. In *IJCAI 2009*, 1586-1591.
- Wang, L., and Cardie, C. 2013. Domain-Independent Abstract Generation for Focused Meeting Summarization. In *ACL 2013*, 1395-1405.
- Zhughe, H. 2009. Communities and Emerging Semantics in Semantic Link Network: Discovery and Learning, *IEEE Transactions on Knowledge and Data Engineering*, vol.21, no.6, 2009, pp. 785-799.
- Zhughe, H. 2010. Interactive Semantics, *Artificial Intelligence*, 174(2010)190-204.
- Zhughe, H. 2011. Semantic linking through spaces for cyber-physical-socio intelligence: A methodology, *Artificial Intelligence*, 175(2011)988-1019.
- Zhughe, H. 2012. Chapter 2 in *The Knowledge Grid: Toward Cyber-Physical Society*. *World Scientific*.
- Zhughe, H. 2015a. Dimensionality on Summarization, *arXiv:1507.00209 [cs.CL]*, 2 July 2015.
- Zhughe, H. 2015b. Mapping Big Data into Knowledge Space with Cognitive Cyber-Infrastructure, *arXiv:1507.06500*, 24 July 2015.