Ontology based Qualitative Methodology for Chinese Language Analysis

Dong Han
Information Management Institute
University of Neuchâtel
Neuchâtel, Switzerland
dong.han@unine.ch

Kilian Stoffel
Information Management Institute
University of Neuchâtel
Neuchâtel, Switzerland
kilian.stoffel@unine.ch

***Abstract*—With its rich history and the rapid economic development in China, Chinese has becoming a language drawing an increased research interest in different domains. The big difference between Chinese and other languages (mainly the European languages) leads to new challenges regarding existing methodologies. In this paper, we propose an ontology-based qualitative methodology for analyzing the Chinese language, aiming to construct machine readable knowledge from original texts, carrying out mining and reasoning on the data, and providing meaningful inference results. Our approach is exploiting characteristics of the language, which allow us to extract certain elements which are fed to the analytical process. A software prototype has been implemented. Experimental cases are presented and they indicate that the approach can produce reasonable results.**

***Keywords*-Chinese; ontology; qualitative research; data mining;**

I. MOTIVATION

As the development of globalization, the importance to comprehend different languages has risen, and the need to be able to analyze them with the aid of computer systems has become crucial. A large number of research activities and industrial applications are hence focusing on linguistic comparisons. IT companies have been providing a list of services such as Google Translate [1] and Bing Translator [2]. Evident achievements have been reached among languages which share similar structures or originate from the same roots. Whilst most of the Indo-European languages are based on alphabets, which are relatively well adapted for formal representations with logical and algebraic methodologies, some other languages are much more challenging to apply existing methodologies. Chinese, for example, comprises thousands of symbols, which increases the difficulty for automatic processing in a smooth way compared to languages based on alphabets. Chinese has a long history indicated by its maintenance of the vocabulary [3]. One of the key elements of this language is the symbols, which act as an important position in the whole society, and the symbols, especially for its accumulation, are affected by the way of speaking and writing [4]. Furthermore, the grammar and structure of the phrases and sentences are different from most western languages. With the increasing demand for using Chinese in an automated way inside and outside China, significant research topics in analyzing Chinese have emerge from different domains. Most of the approaches developed so far are based on linguistic or literal orientation. The study on how to represent it in the form of formalized knowledge is leaping far behind. Motivated by the facts mentioned above, in this paper, we are going to present an ontology based qualitative methodology, to analyze Chinese. The methodology proposed can easily be extended as the basis to more complex topics derived from the one presented in this paper.

Generally speaking, most Indo-European languages comprise a limited number of letters. Even though the representative forms of these letters are not completely identical, they share very similar features. One of the differences is that Chinese depends on a large number of symbol compositions whilst English modifies the words by adding the endings as the way how it changes the forms [5]. Also Chinese has a large vocabulary for advanced expressions with rich and cultural representations, making it more straightforward since it is based on a written form not affected by the pronunciations [6]. All these statements above are the motivations of this paper focusing on the Chinese language as the research point.

On the other hand from a computer science point of view, classical approaches based on quantitative and statistical studies are no suitable to analyse Chinese. A novel methodology is highly in demand. The methodology proposed in this paper is based on ontologies, an approach for knowledge representation and knowledge management. A large number of advantages of ontologies have been discovered in the past years, especially their hierarchical structure, the capacity of reasoning over them, as well as the machine readability. Ontologies allow for an easy incorporation of existing knowledge into system, building up new knowledge from the analysis and implementing the reasoning procedures. Since all the knowledge is stored in the form of ontologies, it is intuitive for linguistic experts to interact with the system. Meanwhile it is also easy to handle the system with automated methods. Ontologies seem particularly interesting for Chinese analysis in that they provide not only syntax structures, but also the semantics.

II. The approach

Chinese language has a large number of native speakers, indicating its global popularity in the world [7]. According to their features, we have selected a qualitative research approach instead of a quantitative one. The major issue in qualitative research is how to collect, maintain and reason on the data. Among all these options, we choose *grounded theory* [8] as the way how we set up the theory during the entire process: in grounded theory, information is gathered in the form of groups of codes and concepts which are formed on top of the codes; once these concepts are well established, we can group them into categories and build up the theories. The procedures of grounded theory fit well the analytical steps of Chinese as it is a language highly focusing on concepts.

Once we have chosen grounded theory, the challenge is how to formalize the data gather during the construction of a theory in such a way that it can by analyzed in a systematic way. Formal normalization is crucial to clarify domain knowledge in many IT applications [9]. With the field of information technology starting to pay more and more attention on ontology [10], we propose to use them on one hand to integrate the domain knowledge as well as to formalize the representation of the Chinese language. The use of ontologies has furthermore the advantage that they permit formal analysis as well as practical implementations. It will also be possible to extend them to accommodate all future developments.

III. Methodology

A. Research targets

Due to the vast content of the language, the first task is to select some of the interesting subsets. In this paper, we are going to limit ourselves to the following questions:

1) *Interrogation*. In English, an interrogative sentence is normally represented by the inverted order of this sentence whilst in Chinese the order is seldom inverted. It is always following the structure of a declarative sentence. There are several approaches to propose an interrogative sentence:
   - *An interrogative particle*. Very frequently, an interrogative particle is placed at the end of the sentence to request an answer.
   - *An interrogative pronoun*. Similar to oral English, interrogative pronoun could be placed inside a sentence following a declarative order.
   - *To interrogate the verb*. Another way to express the questions is to interrogate the verb, such as "you *like or not* this schedule?"
2) *Tense*. Tense is probably one of the most important grammatical points in all the languages. In Chinese, however, the infinitive form of verbs is always kept, extended by different adverbs. For instance, "*Yesterday, I going shopping already*[1]". Instead of changing from "*going*" to "*went*", the adverbs "*yesterday*" and "*already*" indicate the past tense. In other words, there is no transformation of the verbs; only the combinations of verbs and adverbs are used.

In this paper, we will address these two issues because they are very typical for the problems we are facing while trying to automatically process the Chinese language and furthermore it is not possible to borrow methods developed for other languages as they usually do not share these problems. Once these questions are studied, other open questions could be answered in a straightforward way.

B. Paradigm Modeling

After some topics have been selected as the targets, the next step is to build up a data paradigm which will be employed to dissect different sentences in Chinese. The paradigm we established contains the following elements:

$$S = \{\phi, \eta, \gamma, \nu, \theta, \rho\} \tag{1}$$

In this model, γ is the adverb before the verb; η and θ represent receptively subject and object in each sentence while the verb is denoted as ν. At the beginning and the end of a sentence, we add specially two elements named ϕ and ρ as the grammar particles. They will play an important role in our analysis of the sentences.

This paradigm has several advantages: first, it is intuitive to link it to other languages. A reader with some knowledge of the English grammar can understand it without too much effort. Second, it covers a large proportion of Chinese sentences. Third and maybe the most important, it provides a mechanism to make these sentences machine processable. We are certainly aware of the fact that it is not practical to make use of one model for all the Chinese sentences, but our objective is to set up a first model to facilitate the system process and also to bridge the gap between Chinese and English. This model has a clear meaning for information retrieval and is also easily extendable for further developments.

C. Data Selection

On top of the model set up in the previous step, some refined data is supposed to be obtained to train the model. Several criteria for the data selection should be considered:

1) *Appropriate Chinese version.* Considering the long history of the Chinese language, there are several groups of versions some of which can date back two thousand years. We would like to choose the modern

[1]In this paper, we use non-grammatical English sentences as the direct translation from Chinese.

Mandarin; the most pervasive way people speak and write today, as our input data.

2) *Straightforward expressive ways.* There are a very large number of idioms in Chinese, most of them for historical reasons. These idioms are usually very short but equivalent to long sentences. Furthermore these idioms are not self-contained. External knowledge will be necessary to interpret them. For instance, "*Wei wei jiu zhao*" (to surround Country Wei in order to rescue Country Zhao) means if your allied member A is under attack of B, you could consider attacking B's capital as a roundabout approach rather than save A directly. Since there are no obvious structures in these idioms, we do not take them as the input data.

Based on these ideas above, we have established some typical sentences and phrases as the input data and try to analyze them in a systematic way.

D. Feature Filtering

In the next step, we are interested in extracting some key features from the input data based on our paradigm. Several methodologies of data mining have been used in this step, particularly decision tree. Initially, a *bag of keywords* has been set up as a *hash map* with a list of *keys* and *values*. For example, just = *zhengzai*, past = *zuotian*, future = *jianglai*, etc. The bag of keywords is based on external knowledge but the size is very limited. Next, an input table is set up coherent with the paradigm in (1). Here two steps are necessary to convert Table I[2] to a machine readable form: first, iterate the sentences and whenever the keywords are discovered, they are replaced by their keys in the hash map. Once this step is accomplished, we replace the other words by randomized values. Table I illustrates the data in the form of the paradigm we built as well as the tense (S_t) and tone (S_q) of the sentences.

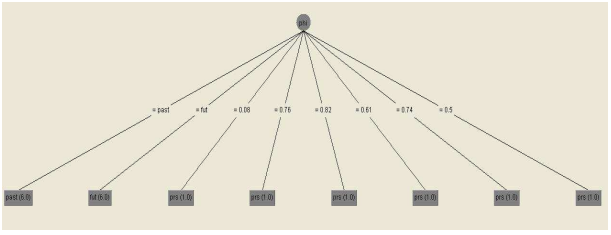

Figure 1. Decision tree 1

Then we used Weka [11], an open source software for data mining to detect the hidden rules. From the decision trees of Figure 1 and Figure 2, it is easily to draw the conclusion that the features of a sentence are closely related to the bag of keywords. These are exactly the features we would like to filter from the data. For example:

$$(\phi = \text{"}past\text{"}) \Rightarrow (s \in past)$$

$$(\phi = \text{"}future\text{"}) \Rightarrow (s \in future)$$

$$(\rho = \text{"}ma\text{"}) \Rightarrow (s \in question)$$

[2]"*le*" means *already* and "*ma*" means *whether/right*(to confirm) in Chinese

ϕ	η	γ	ν	θ	ρ	S_t	S_q
fut	0.16	0.97	0.8	0.11	ma	fut	que
fut	0.16	0.97	0.8	0.11	0.17	fut	sta
past	0.19	0.03	0.11	0.72	le	past	sta
past	0.66	0.71	0.72	0.83	ma	past	que
fut	0.34	0.51	0.74	0.29	ma	fut	que
0.06	0.86	just	0.77	0.89	ma	prs	que
past	0.35	0.81	0.55	0.59	ma	past	que
past	0.55	0.34	0.39	0.45	le	past	sta
0.08	0.66	just	0.58	0.81	0.25	prs	sta
past	0.66	0.5	0.88	0.17	le	past	sta
fut	0.82	0.03	0.14	0.65	ma	fut	que
fut	0.82	0.03	0.14	0.65	0.64	fut	sta
0.76	0.28	just	0.25	0.33	0.62	prs	sta
0.37	0.65	just	0.5	0.66	ma	prs	que
fut	0.34	0.51	0.74	0.29	0.5	fut	sta
0.82	0.27	just	0.68	0.87	0.7	prs	sta
0.66	0.76	just	0.19	0.81	ma	prs	que
past	0.32	0.29	0.48	0.85	ma	past	que

Table I
COMPARISON TABLE

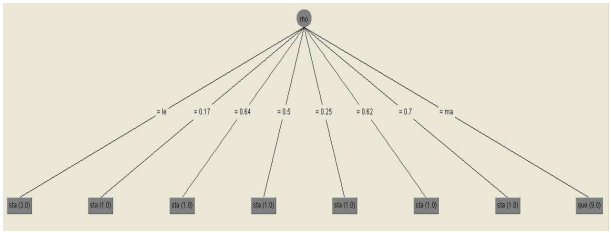

Figure 2. Decision tree 2

E. Ontology Establishment

Based on the features filtered, a group of files in the form of ontologies are created in order to support the knowledge storage and representation. All the ontologies are divided into three categories:

1) *Instance ontologies*. Instance ontologies, as shown in Figure 3, keep all the data from the original sentences. All articles, texts and sentences, can be converted directly into an instance ontology to formalize the data. This type of ontology is usable not only for Chinese, but all the languages.
2) *Speech ontologies*. Speech ontologies are the lists of parts of speech of the words. In Figure 4, for instance, when we see the word "red", we are aware of the fact that it is an adjective. The main purpose of speech ontologies is not to translate the sentences, but to generalize them to our paradigm for the sake of further analysis.

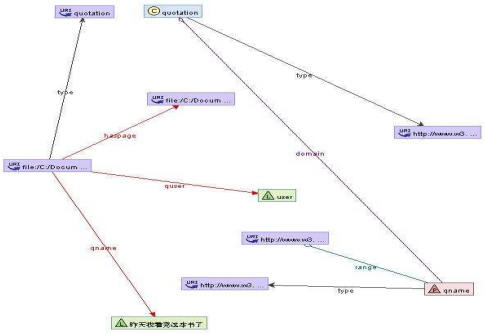

Figure 3. Instance ontologies

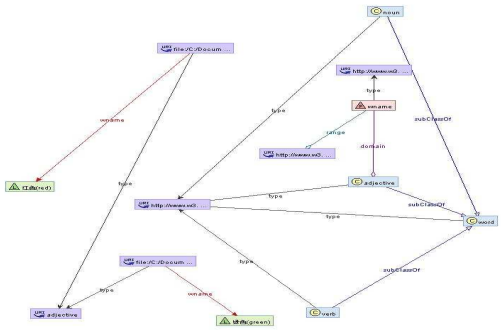

Figure 4. Speech ontologies

3) *Auxiliary ontologies*. Auxiliary ontologies maintain the bag of keywords as shown in Figure 5. These keywords highly influence the grammar of the language. Two similar sentences with slightly different keywords will be different for their tense, tone, or meaning.

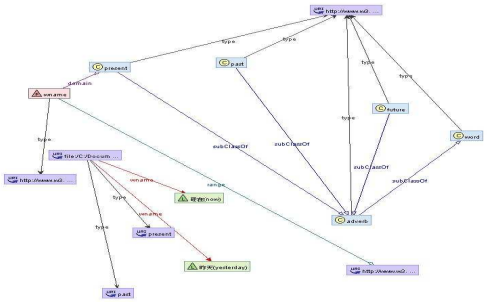

Figure 5. Auxiliary ontologies

All these three categories follow coherently the idea of the *grounded theory*, from the original data, codes to categories and theories. The design principle above has the advantage that it separates the content and the syntax of the language. Later in case some rules or versions are changed, we only need to modify the auxiliary ontologies to form the new versions.

F. Feature Extract Algorithm

The feature extraction algorithm is characterized based on the ontologies designed. The algorithm aims to extract the features from Chinese sentences. The basic idea is to iterate over the original texts, try to find the keywords and their roots in the auxiliary ontologies and generalize other non-keywords by using speech ontologies. The input of the algorithm are the three types of ontologies - instance ontologies (O_i), speech ontologies (O_s) and auxiliary ontologies (O_a). The output will be the feature m (tense, tone, etc) of the sentence q.

Algorithm 1 Extract features

Require: O_i, O_s and O_a as the input, OT as the output
Ensure:
  $O_{temp} \leftarrow O_i$
  while O_{temp} has next **do**
    $q \leftarrow O_{temp}.next$
    while q has next **do**
      $w \leftarrow q.next$
      if $\exists \tau, \tau \in O_s \bigwedge \tau \in w$ **then**
        $w = \tau.getRoot(O_s)$
      end if
      if $\exists \epsilon, \epsilon \in O_a \bigwedge \epsilon \in w$ **then**
        $m = \epsilon.getRoot(O_a).getKey$
        $t = \epsilon.getRoot(O_a).getValue$
        if $OT_{q,m}$ is $null$ **then**
          $OT_{q,m} = t$
        end if
      end if
    end while
  end while

IV. Implementation

In order to provide a working platform, and to conduct our tests, we have developed a software prototype for qualitative studies based on texts and documents. The prototype, based on ICEPdf [12], has multiple functionalities in order to support the analysis in a wide range of domains, such as sustainability and environmental research [13], text recognition and language analysis. The main architecture is presented in Figure 6:

A. Text Annotation

The input of our system is a PDF file. In case of Word, HTML, or other textual files, they can be converted to PDF files. An illustration of the interface of the system is given in the screenshot in Figure 7. The following functionalities are up to this point supported by the system:

- Read a PDF file, select different pages, and zoom into different views. Print the file and extract all the texts.

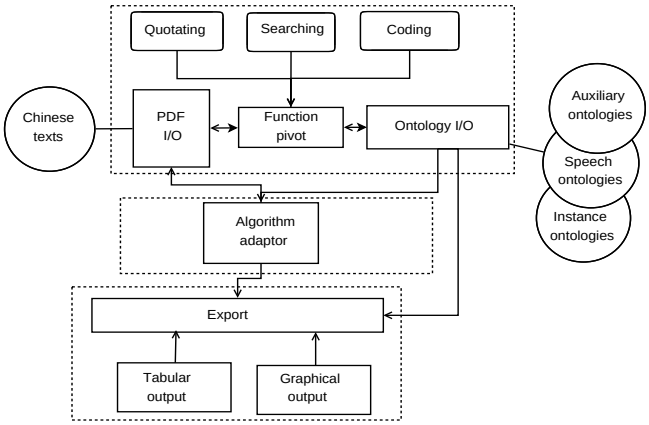

Figure 6. System architecture

- Select a certain part of the texts which a user is interested in as a quotation. Highlight the text parts in the PDF file which have been selected as a quotation.
- Give codes (labels, tags) to the quotations. Search key words from the texts and search codes. Set up code families based on their innate relationships.
- Export the analytical results. Save these results as CSV/Excel files. Export graphical representation in the form of plots. Get proposition for the codes to give to a quotation.

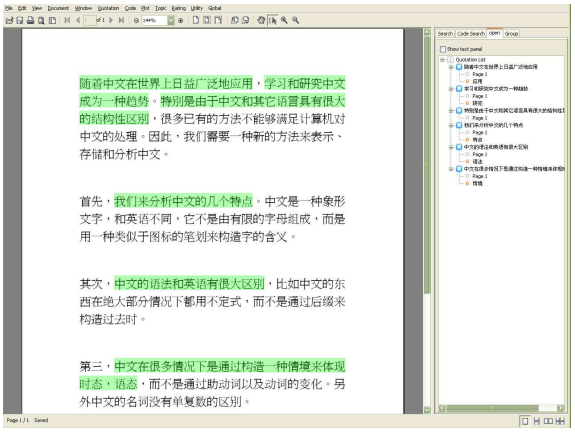

Figure 7. System screenshot

B. Language Analysis

Inherited from the methodology proposed and based on the annotated texts, the prototype contains a couple of steps for analyzing the language. First, the textual input (the original files, quotations, and codes) are delivered to the analysis module. Then the system starts to load the *auxiliary ontologies* to execute the extraction algorithm. Once some results have been produced, it will keep them in memory and push them to the export module for visualization and persistence. For example, a sentence, which is represented as a quotation in our system, will be labeled with a code ”past” if the system discovers that it has past tense.

C. Output and Export

Besides, a list of output and export utilities are also supported:

- All the quotations and codes selected.
- The frequency of codes and code pairs
- The distribution of the codes in terms of pages
- The distribution of the quotation size in terms of pages. As Figure 8,the X-axis represents the pages and the Y-axis stands for the text size selected as quotations in each page.

All the functionalities describe above are obtained by querying the ontologies using XPath and can be saved as Excel/CSV spreadsheets for further analysis.

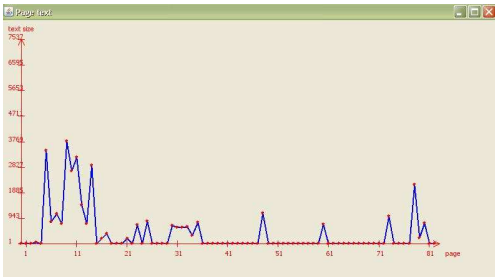

Figure 8. An example of output

V. Experiment and Discussion

In order to test our proposed methodology, we have designed an experiment with ten sentences (in Figure 9) as the input. The objective is to distinguish their tense and tone.

n	Sentences
1	昨天，我去商场了。(Yesterday, he go shops already)
2	他去上课了。(He go to the course already)
3	他去比赛了吗？(You go to the match already right)
4	这本书好不好看？(This book is very interesting or not)
5	我正在讨论问题。(We just discuss questions)
6	每天他都跑步。(Every day, he go running)
7	他跑步去了？(He go running already)
8	你要去哪里？(He want go where)
9	我准备看这部电影 (I plan watch this movie)
10	上次比赛以后，我很累 (After last match, I be tired)

Figure 9. Input sentences

In Table II, we can see that we were able to extract the features from most of the sentences. The left columns stand for the tense and tone in reality whilst the right columns indicate the results from our proposed approach. The results

n	tense_1	tone_1	tense_2	tone_2
1	past	sta	past	sta
2	past	sta	past	sta
3	past	que	past	que
4	prs	que	prs	sta
5	prs	sta	prs	sta
6	prs	sta	0	sta
7	past	que	past	sta
8	fut	que	0	que
9	fut	sta	0	sta
10	past	sta	0	sta

Table II
COMPARISON TABLE

with underscores are the ones with mistakes. For past tense, it is relatively straightforward whilst it is more challenging for future tense. The reason is that future tense in Chinese is always formed with a verb representing behaviors in the future - for example, "I *plan* do something" or "I *want* do something". Even though our algorithm works, auxiliary ontologies with larger vocabularies are needed.

The limitation of the methodology proposed here mainly results from the training data. As the Chinese language has a very large vocabulary, it is not easy to filter the features and thus to build up the ontologies containing a large proportion of these words. This is especially difficult for newspapers, fictions and even scientific documents, as a high amount of rhetoric is used. Besides, the sentences involved do not follow standard grammar; instead, they are creating a scenario for the readers to understand. For example, in an article about the bicycle competition in Beijing [14], only the first sentences mention the auxiliary words concerning the tense. From the middle part of the article, there is no obvious indication about the tense from the verbs. The readers have to do the reasoning by themselves from the adjacent phrases in order to follow the sequence of events. In this case, more advanced approaches are needed to inference the order of events.

VI. Conclusion and Future Work

In this paper, an ontology based methodology is proposed to analyze the Chinese language. Designed in a qualitative way, this approach focuses on a series of innate features of the language, tries to set up a paradigm to model typical sentences and then uses some data mining methods to filter their features. Ontologies are created to represent the original data and the features to support the algorithms developed. We have successfully implemented a prototype for the presented methodology. Furthermore, the experiment carried out indicates that the approach proposed has reasonable results on the test data and also reveals promising research potential.

In the future, our work will focus on two aspects: first, we would like to start to analyze compound sentences which are far more challenging than the sentences used in this paper. Second, irregular sentences have to be integrated into our research scope.

Acknowledgements

This work is supported by Swiss National Science Foundation(SNSF) project *"Formal Modelling of Qualitative Case Studies: An Application in Environmental Management*" (Project No. CR21I2_132089/1).

References

[1] Google, "Google Translate," Website, 2011, http://translate.google.com/.

[2] Microsoft, "Bing Translator," Website, 2011, http://www.microsofttranslator.com/.

[3] W.-L. Soong, "Modeling Presence and Absence in a Few Chinese Semantic Primes," Website, 2011, http://www.ut.ee/SOSE/conference/summer_school/2011/papers/soong_TSSS2011.pdf.

[4] The British Museum, "Chinese symbols," Website, 2011, www.britishmuseum.org/pdf/Chinese_symbols_1109.pdf.

[5] A. Burk, C. Coleman, C. Wimberly, and J. Zapata, "The Chinese Language Manual," Website, 2008, http://languagemanuals.weebly.com/uploads/4/8/5/3/4853169/chinesemanual.pdf.

[6] P. Rouzer, *A New Practical Primer of Literary Chinese*. Harvard University Asia Center, 2007.

[7] Wikipedia, "Chinese language," Website, 2011, http://en.wikipedia.org/wiki/Chinese_language.

[8] B. Glaser and A. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Aldine, 1967.

[9] J. Bouaud, B. Bachimont, J. Charlet, and P. Zweigenbaum, "Methodological Principles for Structuring an "Ontology"," in *Proceedings of Workshop "Basic Ontological Issues in Knowledge Sharing", IJCAI 1995*, 1995.

[10] N. Guarino, "Formal Ontology and Information Systems," in *Proceedings of FOIS*, 1998.

[11] University of Waikato, "Weka 3 - Data Mining with Open Source Machine Learning Software," Website, 2011, http://www.cs.waikato.ac.nz/ml/weka/.

[12] ICEPdf, "ICEPdf Homepage," Website, 2011, http://www.icepdf.org/.

[13] D. Han and K. Stoffel, "Ontology based Qualitative Case Studies for Sustainability Research," in *Proceedings of Workshop "A.I. for Intelligent Planet", IJCAI 2011*, 2011.

[14] Beijing Daily, "The first Beijing Bicycle Match will be held, with 260 bus routines modified." Website, 2011, http://www.bjd.com.cn/10bjxw/shenghuo/201110/04/t20111004_1133447.html.