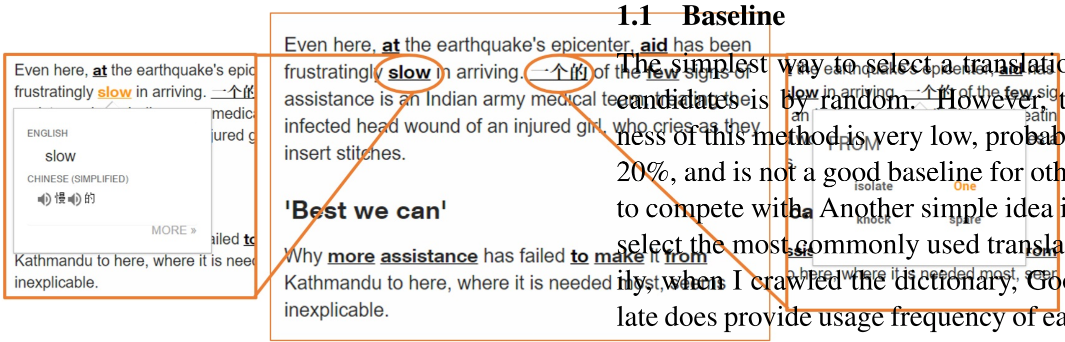
Learning Second Languages from News Websites

Abstract

Learning a second language is difficult and requires constant revision and immersion. Fortunately, many of us read news online everyday. In this paper, we propose a web browser extension that allows readers to learn a second language vocabulary while reading news online. To this end, we propose algorithms to disambiguate word sense and translate the words in new articles properly to target langauge. We find a machine translation based method significantly betters baselines in both coverage and accuracy. We also propose techniques for generating appropriate distractors for multiple-choice word mastery quizzes for assessing language learners. We conducted a user survey to evaluate our system.



Even here, at the earthquake's epicenter, aid has been frustratingly slow in arriving. 一个的 of the few signs of assistance is an Indian army medical team tending to the infected head wound of an injured girl, who cries as they insert stitches.

'Best we can'

Why more assistance has failed to make it from Kathmandu to here, where it is needed most, seems inexplicable.

Figure 1: ....

1 Word Sense Disambiguation System

As we all know, one word may have multiple translations in another language, and our extension is expected to select the most appropriate one based on the context.We call such translation selection as cross-lingual word sense disambiguation (WSD).

In this following, I describe four approaches that I have tried to accomplish WSD system, which is also my main progress in the second semester. The four approaches are:

- Frequency based: always selecting the most frequent translation (the baseline),
- Part-of-Speech Tag based: selecting the translation based on the Part-of-Speech Tag of the English word
- Translation based: Selecting the translation based on the result from existing Machine Translation systems
- Category based: Selecting the translation based on the category of the news article

1.1 Baseline

The simplest way to select a translation from the candidates is by random. However, the correctness of this method is very low, probably less than 20%, and is not a good baseline for other methods to compete with. Another simple idea is to always select the most commonly used translation. Luckily, when I crawled the dictionary, Google Translate does provide usage frequency of each Chinese Translation. This turns out to be a much better result, and thus serves as a fair baseline method.

1.2 Part-of-Speech Tagger

As we all know, many English words have more than one Part-of-Speech (POS) tags and their Chinese translations in different POS may differ a lot. For example, the word “book” has two POS tags, noun and verb. If it is used as a noun, mostly

Table 1: Example input/output of WSD.

English Sentence	Word	Dictionary	Baseline	Category	Bing	Bing+	Bing++
... treating me like family ...	like	verb : 喜欢, 爱... ... preposition：好像, 好比 ...	喜欢	好像			
... painting a picture of urban street life ...	picture	... 相, 影, 影片(entertainment), 帧, 想象, 画 ...		影片			
... pistol a pump shotgun ...	pump	verb:抽, 抽水, 打气, 唧, 唧筒, 套 noun:抽水机, 唧筒			唧筒		
... have made it into the worlds top 40 clubs ...	top	顶部, 顶端, 顶, 颠, 盖, 极 ...	顶部		顶	顶级	
state department spokeswoman ...	state	...陈, 陈说, 称, 称述, 发表, 发言...			发言	发言人	国家

it means a handwritten or printed work of fiction or nonfiction, which should be translated as “书”, and mostly means to reserve if used as a verb, which should be translated as “预定”. Therefore, getting the POS tag of the English word might help us identify its sense or the Chinese translation. We decide use Stanford Log-linear Part-of-Speech Tagger (**?**).

Firstly, if the word ”like” need to be translated, the algorithm will fetch all the Chinese translations as well as their Part-of-Speech tag from our dictionary. Secondly, the algorithm will send the original English sentence to Part-of-Speech Tagger, which is a Java package and has been wrapped into a server. After the client has got the output from the server, it will fetch the corresponding tag and match it to Part-of-Speech tag based on the guidelines mentioned above. Lastly, it will select the translations based on the POS.

1.3 News Category

The word “interest” have two very different translations when it is used as a noun. One translation is about “the feeling of a person whose attention, concern, or curiosity is particularly engaged by something”, which should be translated as “兴趣”. The other translation is about “a share, right, or title in the ownership of property”, which should be translated as “利息”. It is quite obvious that the second sense is mostly used in financial related topics. Therefore, analysing the category of the original article and selecting the translation with the same category label might help disambiguate the word meaning.

In Table 1, the second example, word “picture” is the word that need to be translated. Firstly, the algorithm will fetch all the Chinese translations for word “picture” and only the word “影片” has a category “entertainment”. Next, the algorithm will fetch the category of the English news article from the URL, which is also “entertainment”.In this case, the algorithm will use “影片” as the translation for word “picture”. If a few words shares the same category, the algorithm will choose the translation with the highest frequency of use.

1.4 Machine Translation

Since our target is to select the most appropriate translation based on the context, using existing Machine Translation (MT) systems is also a good approach, as all of them will certainly translate words based on the context. After I tried a few on-line or off-line MT systems, We decide to use Bing Translator as our Machine Translation system.

1.4.1 Bing

In Table 1, the thrid example, the original English sentence is “including a 45-caliber pistol a pump shotgun and an ar-15 rifle” and “pump” is the word that we want to translate. Firstly, this algorithm will fetch all the Chinese translations from the database. Next, it will send the original English sentence to Bing Translator using the API provided by Microsoft and get the result that returned from Bing Translator. After that, for each Chinese translation, I will check whether this translation is a substring of the Bing Translator result. If there are a few translations that can match with the Bing Translator result, I will select the longest translation. If there are a few translations with the same length and all of them can match with the Bing Translator result, I will select the translation with the highest frequency of use. In this example, both “唧” and “唧筒” are the substrings of Bing Translator result. As “唧筒” have two characters and “唧” only have one character, this algorithm will take “唧筒” as the final result.

1.4.2 Bing+

Bing approach is not perfect. The results that generated by Bing approach is limited by the covearge of our dictionary size. In Table 1, the fourth example is the approach of using Bing Translator together with Stanford Word Segmenter, and I would like to use Bing+ to represent this algorithm. The Bing approach will generate “顶” as the result. After that, our algorithm will send the Chinese sentence returned from Bing Translator to Stanford Word Segmenter. Then, this algorithm will use the segmented word that contains the Bing result as a substring or equals to the Bing result as the final result. In this example, the final result of Bing+ is “顶级” which is the best result that can be generated from the result of Bing Translator and also a result that does not covered by our dictionary.

1.4.3 Bing++

Bing+ approach is not perfect as well. The results from Bing+ approach is highly related to the accuracy of string matching algorithm. If two English words shares very similar translations or if two Chinese words contains the same Chinese charater, Bing+ approach will generate the wrong result and that’s why we need a Word Alignment tool.Bitext word alignment or simply word alignment is the natural language processing task of identifying translation relationships among the words (or more rarely multiword units) in a bitext, resulting in a bipartite graph between the two sides of the bitext, with an arc between two words if and only if they are translations of one another. I use Bing Word Alignment API[1] as our Word Alignment tool. The Bing++ algorithm is basically the approach of using Bing+ approach together with the Microsoft Bing Word Alignment. In Table 1, the fifth example, “state” is the word that need to be translated. The result from Bing+ approach is “发言人”, which is the translation of “spokeswoman”, because the Chinese translation “发言” can be translated from both “state” and “spokeswoman”. Then step five will send the original English sentence to Bing Word Alignment. Now, there will be two final results, one from Bing+ approach and the other one from Bing Word Alignment and the algorithm will choose the correct one from these two results. In this example, “state” will match with “国家” and the algorithm will choose “国家” as the final result as well.

1.5 WSD System

Our Word Sense Disambiguate System can be evaluated from two important aspects: coverage (i.e., is able to return a translation) and accuracy (i.e., the translation is proper). To this end, I manually annotate the ground truth.

Table 2: Coverage for different approaches

	Coverage	Accuracy
Baseline	100%	57.3%
POSTagger	94.5%	55.2%
News Category	2.0%	7.1%
Bing	78.5%	79.8%
Bing+	75.7%	80.9%
Bing++	76.9%	97.4%

Table 2 column two contains the coverage for different approaches. As the algorithm will try to translate some word only if it is covered by

[1] https://msdn.microsoft.com/en-us/library/dn198370.aspx

our dictionary, the coverage for Baseline is always 100%. The coverage for Bing, Bing+, Bing++ and POSTagger are roughly the same and all of them are acceptable. However, the coverage for News Category approach is only 2.0%. One reason is that when I set the threshold for assigning categories for Chinese word, I purposely make it very high to maximize the accuracy. If the accuracy is quite high, which means this approach is quite useful, then I will lower the threshold and find the balance point.

Figure 2 column three contains the accuracy of all the approaches. The last column is the accuracy for News Category approach and it is only 7.1%. As mentioned in above Chapter, since the accuracy is very low, there is no need to lower the threshold and try to allocate more categories for Chinese words. The accuracy for Baseline is 57.3%, which is already a fairly hight accuracy. The accuracy for POSTagger is around 55.2% also, which is a bit lower than our expectation. The accuracy for Bing++ is 97.4% which I think is a very good result and it is already very hard to improve. Therefore, based on my test results, Bing++ is the best approach among these five approaches.