QingheGao
2659024

Assignment 1
NLP Technology

2020/04/11

1 Linguistic Fundamentals

1.1 Linguistic Concepts

Table 1:

Concept	Definition	Examples	NLP relevance	Reading
morpheme	morpheme is ”the smallest meaningful units of language”(Bender, 2013, p.11).	*Apples* consists of apple+s, a root and plural marker.	morpheme can be used to identify the roots of words to reduce ambiguity and also for part-of-speech.	B[6] 7,12
lemma	Lemma is a basic form or a root of words and also can be explained like a dictionary form of a word.	Basic form of *car, cars, car's, cars'* is car. The root of *girl, girls, girl's, girls'* is girl.	Lemma can be used into Lemmatisation to get the root of word and also can be used into word sense disambiguation. It also can support POS-tagging and sentiment analysis.	E[9].4.2
pos-tag	pos-tag is a process that marking up each word to put into corresponding part-of-speech based on context.	*I like hamburgers.* I and hamburgers are nouns, like is verb. And in other context, *Sky like a carpet.* Sky and carpet are nouns, a is an article and like is a preposition.	pos-tag can be used into disambiguation and also can enhance word-based features.(eg: book my hotel, I like this book. Two book are different part-of-speech). And also for sentiment analysis, it is also useful especially for lexicon based method.	E[9] 8.1; JM[12] 8.1, 8.2
constituent	The words are comparatively closer to one another than to other words in a sentence or ”Words within sentences form intermediate groupings”.(Bender, 2013, p.61)	Tom eats [a very delicious pancake]. *a very delicious* is comparatively closer to *pancake*.	Constituent can be used to analyze composition and structure of phrases in sentence, which can be helpful to do analyze the relation between phrases to finish certain task such as reducing ambiguity, getting sequence of sentence, modality and so on.	B[6] 51, JM[12] 12.1 - 12.3
dependency	Dependency is the grammatical relation between words and the relation is binary.	*I saw a girl.* I ← saw → girl. Girl → saw. *saw*'s grammatical relation are *I* and *girl*. And *girl* is grammatically related to *a*.	Dependency can be applied into reduce ambiguity of sentence, and dependency structure can illustrate root node, head of structure and syntactic analysis.	JM[12] 15.1, 15.2

It took me four hours to complete all the stuff and read material.

1.2 Linguistic Analysis

1.Wuhan (the city where the virus originated) is the largest city in Central China, with a population of over 11 million people. 2. The city, on January 23, shut down transport links.

1.2.1 UD POS-tag

For the first sentence:

Words	UD tag	Words	UD tag
Wuhan	Proper nouns(PROPN)	the	Determiners(DET)
city	Nouns(NOUN)	where	Adverb(ADV)
virus	Nouns(NOUN)	originated	Verb(VERB)
is	Verb(VERB)	largest	Adjectives (ADJ)
in	Adposition(ADP)	Central	Noun(NOUN)
China	Proper nouns(PROPN)	with	Adpositions(ADP)
a	Determiners(DET)	population	Nouns(NOUN)
of	Adposition(ADP)	over	Adposition(ADP)
11	numeral(NUM)	million	numeral(NUM)
people	Nouns(NOUN)		

Table 2: UD POS-tag for first sentence

For second sentence:

Words	UD tag	Words	UD tag
the	Determiners(DET)	city	Nouns(NOUN)
on	Adposition(ADP)	January	Proper nouns(PROPN)
23	numeral(NUM)	shut	Verb(VERB)
down	Particle(PRT)	transport	Nouns(NOUN)
links	Nouns(NOUN)		

Table 3: UD POS-tag for second sentence

1.2.2 Morphological analysis

Morphological analysis:

Virus: Noun. This is a noun in this sentence and in this context it mean coronavirus.

Originated: originate(verb)+d(passive marker). And in this sentence originated is an adjective of *virus* and means come from.

Largest is combined by large(adjective)+st(marker of superlative adjective). Largest in the sentence is a adjective and semantic effect is superlative, which can emphasize the size of city and also emphasize the seriousness of virus.

Links is combined by link(noun)+ s(plural marker). Links is noun in this sentence, which means distributing center and links phrase with transport in this sentence.

1.2.3

This bridge links to a restaurant. In this sentence, links is combined by link(verb)+ s(agreement marker). And it means ”connect” and is different from the ”links” in previous sentence.

1.2.4

For first sentence, the root should be *is* not the *city*. Thus the line pointed from *Wuhan* to *city* should point from *is* to *Wuhan*. And the line pointed from *city* to *is* should point from *is* to *city*.

For the second sentence, there are some obvious mistakes. In this sentence, *shut* is the root so a line should point to city with *nsubj* labels. And the line pointed from *city* to first *,* with *punct* should point from *shut* to first *,*.

And this part cost me four hours.

2 NLP tasks: definitions and identifying features

2.1 NLP Tasks: Gold Data

2.1.1

2.1.2 BIO-style

The BIO-style:

(a) The third mate was Flask, a native of Tisbury, in Martha’s Vineyard.
O O O O B-PERSON O O O O B-LOC O O I-Loc I-Loc.

(b) Its official Nintendo announced today that they will release the Nintendo 3DS in north America march 27.

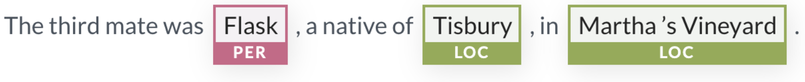
O O B-ORG O B-DATE O O O O B-PRODUCT I-PRODUCT O B-LOC I-LOC B-DATE I-DATE.

(c) Jessica Reif, a media analyst at Merrill Lynch & Co., said, “If they can get up and running with exclusive programming within six months, it doesn’t set the venture back that far.”

B-PERSON I-PERSON O O O O O B-ORG I-ORG I-ORG I-ORG O O O O O O O O O O O O O O O O O O O O O O O O O O O

2.1.3 Allen NLP system

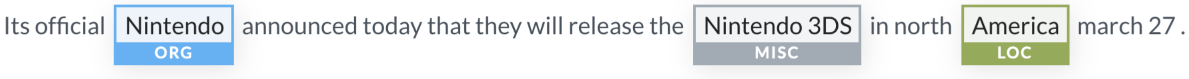
First:



The third mate was Flask , a native of Tisbury , in Martha 's Vineyard .

Figure 1: First sentenc result of Allen NLP system

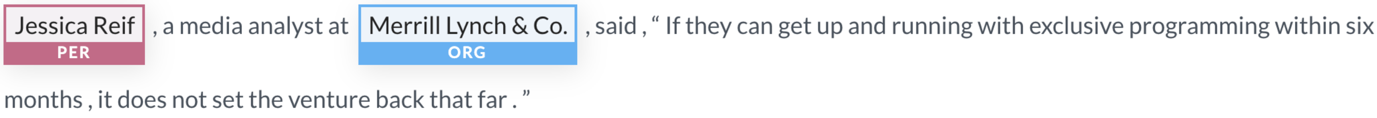
Second:



Its official Nintendo announced today that they will release the Nintendo 3DS in north America march 27 .

Figure 2: Second sentenc result of Allen NLP system

Third:



Jessica Reif , a media analyst at Merrill Lynch & Co. , said , “ If they can get up and running with exclusive programming within six months , it does not set the venture back that far . ”

Figure 3: Third sentence result of Allen NLP system

First result and third match matches my result.

Second result is different from my result. Because *Nintendo 3D* should be *PRODUCT* and *march 27* should be *DATE*. For *march 27* I think it is because of the capitalization of march, the Allen should also include lower case. And for product, I think maybe should update lexicon.

And this section took me three hours

2.2 NLP Tasks: Features

2.2.1 Sentiment Analysis

For sentiment analysis

Features: Feature of single words: 1.Expanding contraction(find symbol like ’ and also can detect some words like *n’t, ’ll,’d*) 2. Tokenization(Just split words by space) 3. Lemmatisation(detect begin of words and match with lexicon) 4.Remove stop-words(remove words with only have one or two letters) 5. Lower the words(detect capital letter) 6.POS-Tagging 7. Negations Handling(not, noun). Also need feature of phrases. 8. constituent. Also feature of Words in Context.10. dependency(word order)

Design choices: Firstly, Expanding contraction is necessary. For example *doesn’t* need to be expanded into *does not*. Otherwise the result will not be accurate. Secondly tokenization is used to split the sentence or paragraphs into every single words. Then lemmatisation can be used to get the roots of each words, which is helpful to finish the sentiment analysis for lexicon methods or other methods. Furthermore, some stop-words need to be removed such as *a, an, the, on,. . . .* Because most of time these stop-words are not useful to do the sentiment analysis. The we need to lower the words. For example, changing *Changing* into *changing*, because when using lexicon method the words need to have lower format. Then the next step is POS-Tagging to get each word’s part-of-speech, which can help to quickly find the scores if using lexicon methods and also is helpful to reduce ambiguation. We also need to handle the negations such as *not, none, ...* Because this negations can simply change the whole attitude into different attitude. For example *She likes, she doesn’t like*. For feature of phrases, we also need to do analyze by consituent and dependency. For example, *I don’t like this* *movie, but he does.* The finally for lexicon method, we can match all the words with given lexicon to get the final score.

This part the information is mainly form [10]

2.2.2 Part-of-speech tagging

For Part-of-speech tagging, lexicon method is mainly discussed in this part.

Features: Feature of words: 1.Expanding contraction 2. Tokenization for example for some words end with *tion*, we know this is Noun. Feature of words in context: 4.dependency(word order)

Design choices: For POS-tagging, the most important is feature of words. This could be achieved by Tokenization, lexicon besed method, or distributional Information to get every single word.For example for some words end with *ous, tive,*, this is feature of adjective. And for feature of words in context, it is also important for example *transport links(Noun), links(Verb) to different place*. The *links* has different meaning for different context. So we need to do analysis such as dependency to get the position or scale of the words in the context. Then we can match the words with lexicon to get the part-of-speech.

2.2.3 Named Entity Recognition

For Named Entity Recognition:

Features: Feature of words, 1. Tokenization 2.POS-tagging 3.Gazetteer(for example, the words with Capital beginning). 4. BIO style. Feature of words in context: 5.dependency(word order).

Design choice: For name entity recognition, first we should should split the words with several features by Tokenization, POS-tagging, Gazetteer and so on. For example, capitalization(Wuhan, Netherlands), medicine(begin with oxa). And we also can POS-tagging to select the part-of-speech. For example most of noun words can be selected and delete other type part-of-speech words. Another example we can set if we found *in, at,*, we can probably know the location words follow these words. And in this way we can extract the location words. Also we should consider the words in context. We can use dependency analysis to get actual meaning in certain context. For example *Central China*, they are both noun.

This part the information is mainly form [11].

3 Crosslinguistic variation

3.1 Linguistic concepts across languages

3.1.1

Classification of writing system is quite diverse, and the broad categories are alphabetic writing system, syllabic writing system, and logographic writing system.

3.1.2

For English, plural marker: apple*s*,orange*s*. Agreement marker: she sing*s*. *s* means subject is first person plural.

For Portuguese adjectives there are gender maker: male: louc**o**. female:louc**a**. [7]

Eastern Pomo has evidential maker to describe the evidence of the message. puna kayan-*a!*(I saw the fire in person). puna kayan-*e!*(I didn't see it, but I feel there was a fire). puna kayan-*ye!*(I didn't see it, but I heard there was a fire)puna kayan-pek!(I didn't see it, but there is a evidence can guess a fire happened. [1]

3.1.3

Morphologically rich means expressing different meaning or grammatical relations can be achieved by changing words. In other words a sea of noun forms or verb forms can be derived from only one word root.

For example: Hebrew. About 120 inflection of each verb-root + prefixes and suffixes. Each noun has about 75 forms.[8]

3.1.4

The main difference of inflectional and derivational morphology is if it will change part-of-speech of words. Inflectional morphology never change part-of-speech of words. For example, +*s*: small(adj) and smaller(adj). While for derivational morphology it usually change part-of-speech of words for example, +*er*: sing(verb) and singer(noun).

For Chinese, inflectional morphology can be achieved by plus *men* to describe plural format: $w\check{o}$(I, Personal Pronoun), $w\check{o}men$(us , Personal Pronoun). For derivational morphology adding *ren* can change adjective into noun. For example: $m\check{e}i$(beautiful, adjective), $m\check{e}iren$(beauty, noun).[4]

3.1.5

Agreement is the form of one word's changes dependently on the form of another word or phrase in a sentence. For example, *I am, he is*. And there is no *I is, he am*. [3]

3.1.6

Grammatical functions basically is defined by word order in the sentence. Subject is noun or pronoun to do the action. And predicate is usually a verb to tell the specific action. Object is the noun which accepts the action performed by subject. And modifier give detailed information.To distinguish subject, subject usually appear to the left of verb. And words order is alo helpful in general.

Eg: He give her a interesting book. He: subject, give: predicate, her: object, a interesting book: modifier. And also the object can be defined by direct object and indirect object. [2]

3.1.7

Word order can not have same significantly impact on many non-projective structures language. It is stable for word order in English, thus the results can be given by analyzing word order. While for languages which have many non-projective structures it is hard to get the conclusion simply from the word order because the word order is flexible.

It took me four hours to finish this.

3.2 Dealing with different languages

In this part, Chinese was used to explore sentiment analysis, part-of-speech tagging, named entity recognition.

For sentiment analysis, features for words: tokenization, Remove stop-words,POS-Tagging, Negations Handling(not, noun). Also features of phrases and words in context.

first step is Chinese tokenization. Chinese is different English, and there is no space between the characters. And one method is importing a dictionary which include all characters and scanning the input to match the words in dictionary to do the tokenization. And phrase non-projective dependency should be done. Because for example *hao* mean *good*, *cha* means *bad*. But *hao cha* means *extremely bad*. Thus, feature of distance of negative and positive should be considered. Then after tokenization, deleting useless words in sentences for example *shi* mean *is* in English. Then for POS-tagging using lexicon to detemine the part-of-speech. Furthermore, the words can match lexicon which includes all the negative and positive words to finish the sentiment analysis.

For part-of-speech tagging, features for words: tokenization, also features of phrases and words in context: constituent and non-projective dependency(words order in compounds).

first step is still tokenization but in this step but it is difficult to Chinese to do tokenization. Because the different segment of sentence could have greatly different meaning. For example, *ta shi ge mei ren*. If *mei* is seprated from *ren*, *mei* is adjective and it means beautiful. While if we combine *mei ren*, this phrase is noun and means beauty. And the meaning of this sentence is *she is beauty*. Thus, constituent and non-projective dependency need to be done for part-of-speech tagging. Constituent and non-projective dependency should be used to analyze compounds of sentence.

For named entity recognition, again first step is still tokenization. In this way the features can be some indicator words for example if *zongli(means Prime Minister)* is in a sentence, and we can know there is a name in this sentence. Thus, we can use a lexicon to store these indicator words. And also using POS-tagging to get part-of-speech, then deleting irrelevant part-of-speech for example adjective words.

This took me four hours to finish.

4 NLP tasks and Gold Data

4.1 Gold data and Evaluation

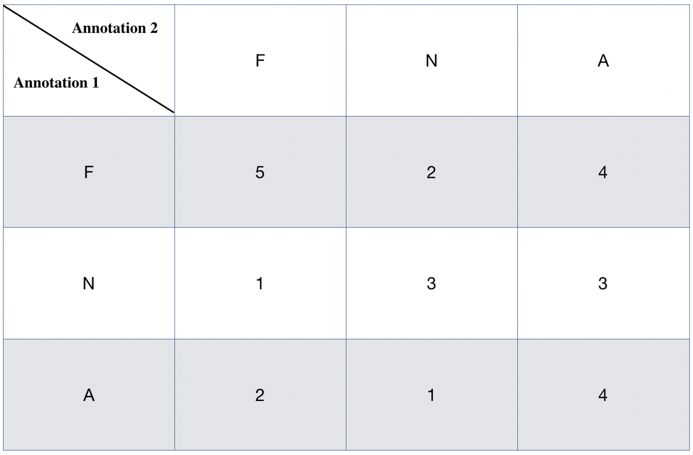
4.1.1 Creating a Gold Standard

Inter-annotator agreement of two annotation was calculated in this part, which is measured by the Cohen's Kappa which can determine the accuracy of classifier. The main idea is using confusion matrix(4) to calculate the Cohen's Kappa. And the result is calculated by equation 1

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where $p_0 = \frac{FF+NN+AA}{total}$ is same prediction of two annotators, $p_e = \frac{a_1b_1+a_2b_2+a_3b_3}{total^2}$, And a_n, b_n are sum of numbers for each label.

And the result of two annotations is 0.222, which has low agreement accuracy. That's because these annotations were made by student, and person's evaluation is subjective.



Annotation 2 / Annotation 1	F	N	A
F	5	2	4
N	1	3	3
A	2	1	4

Figure 4: Confusion matrix of two annotations

4.1.2 small experiment

In this part we did small sentiment analysis of movie review. The 2000 reviews are from Pang and Lee movie review data. First we split the 2000 reviews into 1600 train set and 400 test set. We use two different method to compare the accuracy and F1 measures.

First, lexicon method has been used. In this part we used SentiWordNet lexicon, which contains part-of-speech, negative scores and positive scores for each words. We firstly expand contraction of review, for example expanding *don't* into *do not*. Then using python function to do the tokenization. After tokenization each words match the words in lexicon to calculate the negative and positive scores. And if score is below 0 the review is negative otherwise it is positive.

Secondly, we use machine learning method. The words vector of train set was created first, and then we used Logistic Regression to train the words vector with corresponding label and created a classifier . Final step is using classifier to predict labels of test set.

The accuracy and f1 measure of lexicon based classifier are 0.69, 0.58. And accuracy and f1 measure of logistic regression are 0.81, 0.81. And firstly the a two-tailed binomial hypothesis test was used to measures the difference in accuracy. Using lexicon based method as base line, the p-value is $1.83 * 10^{-152}$. And we also used proportions , the p-value is also super low, which is $2.6 * 10^{-6}$. It means the accuracy of two classifier have significant statistical difference. That's because lexicon based classifier's result greatly depends on preprocess of data. Some process such as remove stop-words did not include in the lexicon based classifier. Thus, the lexicon based classifier has low accuracy when compared with logistic regression.

2 days

4.2 NLP Tasks and Features (2)

In this part in order to apply hedges, denials, and hypotheticals into the NLP, process of extracting features and design choice should be discuss again. There are two basic steps: firstly the negative words in sentence should be clarified and secondly the scope should be defined in the sentence. For hedges, features of words-Tokenization: words which express subject's belief in the propositional content. For example *perhaps, definitely, have to, may*. can be stored in lexicon. And we also can detect if there is url or citation in the sentence.

For negation and modality, feature of words: Tokenization 1. words which express subject's desires. for example: I *wish* you can have a good time. 2. words which express subject's belief in the propositional content. For example *perhaps, definitely, have to, may*. 3. words with N-grams words. For example *not, none, no*. 4.

words with lexical negations. For example *without, lack of*. Feature of words in context. For example, we can detect the distance between modality and negation.

For design of choice, first expanding contraction should be done for example *doesn't* become *does not*. In this way negative words can fully show in sentence. Then using features mentioned before to do toknaization, and each words using POS-tagging to get part-of-speech. And running through all the words to match with lexicon to get all negation and modality in the sentence. Next step is to get the scope for each negative word. For the scope define we need use gazetteer or BIO-style to get the order of the words to math the negation and modality words in a certain scope. Finally we can use syntax and semantic analysis to get conventional sequence. [5]

It took me five hours.

References

[1] Evidentiality. `https://en.wikipedia.org/wiki/Evidentiality`.

[2] introduction-to-english-grammar-and-mechanics. `https://courses.lumenlearning.com/boundless-writing/chapter/introduction-to-english-grammar-and-mechanics/`.

[3] oxfordbibliographies. `https://www-oxfordbibliographies-com.vu-nl.idm.oclc.org/view/document/obo-9780199772810/obo-9780199772810-0118.xml`.

[4] Inflectional and derivational morphemes. `https://semanticsmorphology.weebly.com/inflectional-and-derivational-morphemes.html`, 12 2017. Accessed: 1-12-2019.

[5] Benamara, F., Chardon, B., Mathieu, Y., Popescu, V., and Asher, N. How do negation and modality impact on opinions? In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics* (2012), Association for Computational Linguistics, pp. 10–18.

[6] Bender, E. M. Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax. *Synthesis lectures on human language technologies 6*, 3 (2013), 1–184.

[7] Chahuneau, V., Schlinger, E., Smith, N. A., and Dyer, C. Translating into morphologically rich languages with synthetic phrases. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (2013), pp. 1677–1687.

[8] Cohen, Y., Ben-Simon, A., and Hovav, M. The effect of specific language features on the complexity of systems for automated essay scoring.

[9] Eisenstein, J. Natural language processing, 2018.

[10] Kolchyna, O., Souza, T. T., Treleaven, P., and Aste, T. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955* (2015).

[11] Li, J., Sun, A., Han, J., and Li, C. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[12] Parsing, C. Speech and language processing.