A System for Context-Relevant Term Search

CS 4501 Information Retrieval, Fall 2017

Eric McCord-Snook
University of Virginia
esm7ky@virginia.edu

Daniel Choi
University of Virginia
dc9db@virginia.edu

ABSTRACT

In this paper, we present a new system for providing context-relevant term definition. We show through our examples and experiments how the system can be applied, and our results indicate that the system is objectively successful. We analyze future extensions of this work and improvements to the system that could make it even more effective.

ACM Reference Format:
Eric McCord-Snook and Daniel Choi. 2017. A System for Context-Relevant Term Search: CS 4501 Information Retrieval, Fall 2017. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Our goal for this project was to create a system that can provide an appropriate definition for a search term in context. We define the word "context" to mean the sentence, paragraph, or more general document in which the search term is contained. We reasoned that, in order to help a user understand the meaning of a term in context, we could use the context itself as input to determine which definition makes the most sense. This is the basic premise of our system, and all of the implementation decisions we made keep this goal in mind. The system is meant to handle anything from common homonym disambiguation to complex concept definition in the fields of science, politics, and other niche areas.

Our motivation for this project stems from the common desire to be able to find the correct definition of an unknown word. Since some words are homonyms, i.e. have multiple meanings, a simple dictionary look-up is not good enough for this task. When reading papers written at an advanced level, it is very common to come across terms, concepts, and ideas that we are unfamiliar with. This is an especially prevalent issue for people who have English as a second language. Even looking up words in the dictionary can be insufficient to satisfy their information need, since the multiple definitions provided without proper context can serve to further confuse them; more than one definition could fit, but the resulting context could then take on a whole different meaning.

This problem seems to be one that arises naturally as a result of diversification of the population. As more people from various backgrounds, languages, and cultures join the population, conflicts and troubles born from the different languages are inevitable. Such problems arise anywhere from elementary schools to professional interactions, but are much more common and serious at the professional level of any field where technical terms are used. With terminologies sometimes difficult even for native English speakers, most non-native speakers tend to have trouble or get lost in reading and understanding research papers.

This problem deserves great attention in that it applies to such a large portion of the population today. Since the entire world is connected, wherever we go, it is not a difficult task to find someone whose first language is not English. In such a world, we believe that it is essential to turn our attention to finding an efficient solution for this problem.

In Section 2 of this paper, we go through some of the related work in this area. In Section 3, we discuss our method, and the different working components of the system. In Section 4, we cover our experiment to prove system efficacy and analyze the results. In Section 5, we discuss some of the limitations of the system as it is currently implemented. In Section 6 we discuss possible improvements to the system and how it could be used in the future. We conclude in Section 7.

2 RELATED WORK

Currently, there are several attempts at solutions to the problem we have discussed, with the most general ones being online dictionaries. However, as discussed before, simple dictionary look-up is not enough to solve this problem. A possible solution for people with English as a non-primary language is the set of online translation tools, such as Google Translate. While these are effective to a degree, they fail to understand and choose correctly translated words in the context of the given document a person is reading, and frequently give inappropriate translations as a result. Even with systems that output a list of translations as a result of a query input, it is up to the user to choose which definition is correct, which isn't always easy and doesn't solve the problem at hand. Additionally, when translating a whole sentence or paragraph of text, it can be difficult to figure out which word in the translated output was the original "problem word", leading to confusion in future occurrences of the word. We believe that our novel solution to this problem would be effective in alleviating many of the issues the current tools have.

Several browser extensions already exist which aim to solve a similar problem to the one we've proposed. However, these solutions act more as a direct link to basic dictionary search or Amazon search, again not taking the context into account. Figure 1 shows an example of this type of system in action. Additionally, Apple's MacBook computers come with a "look up" capability built-in. This capability, however, also just does a simple dictionary search for the highlighted term and displays the results nicely. Figure 2 shows the usage of the MacBook "look up" feature, and it is clear that the

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

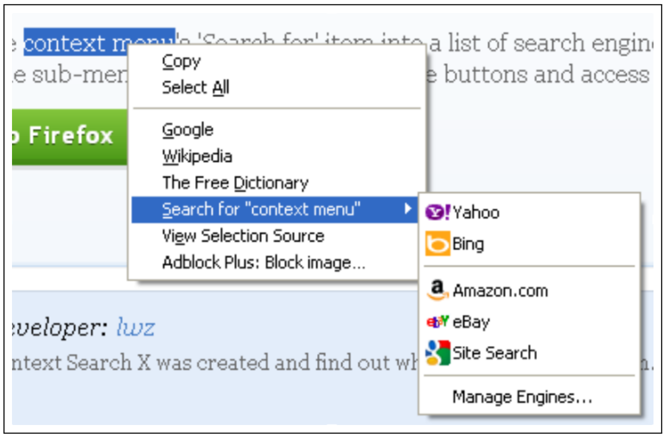

Figure 1: Firefox provides external links to search the web for the search term.

context isn't considered when looking up the definition of the word. Even though the correct result is listed as the second definition, someone who isn't familiar with English would have no way of knowing which definition was correct.

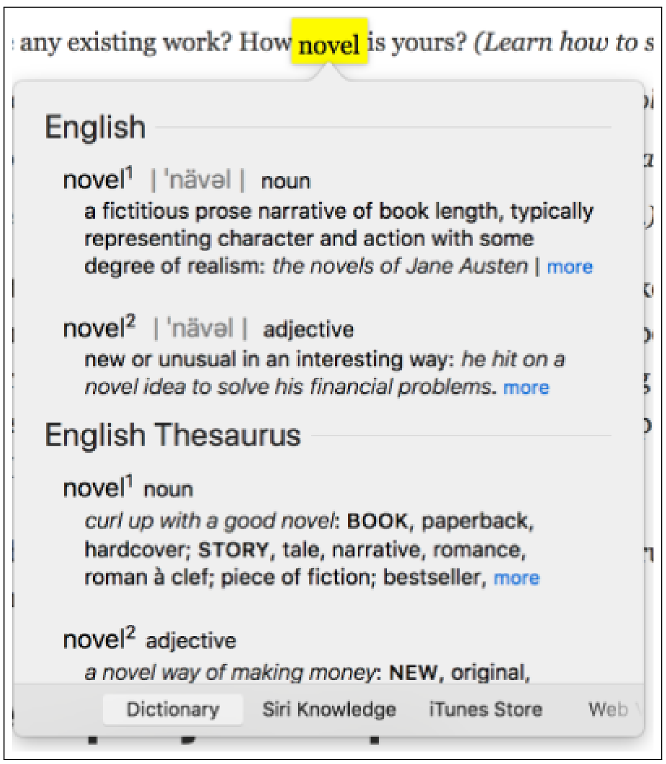

Figure 2: MacBook's "look up" capability gives the incorrect definition of novel in context as its first result.

Overall, no currently existing system implements an actual solution to the problem we've discussed. This further increases the need for a system like the one we've developed.

3 METHOD

In order to build the proposed system, we first needed to define our inputs and outputs, so that we had a clear starting position and end goal to work with. We decided that the inputs to the system would be comprised of the search term and the context for that search term as defined in Section 1. The output of the system was chosen to be a ranked list of possible definitions for the search term in context.

For the system to function as desired, we next needed to devise a method to appropriately predict what the context-relevant definition for the search term would be. Unlike traditional information retrieval systems that find the best match between a query and a collection of indexed documents, however, our system doesn't have a set collection of documents to choose from. Instead, we opted to create our document collection on the fly. For every word, we produce a new collection of documents relative to that word specifically.

Each of the documents that we create represents a single definition for the given search term. We then use Lucene to index these documents and find matches between the input context and the documents in the collection. We choose to use the Jelinek-Mercer smoothed Language Model as our document scoring function, as it gave the best performance on empirical tests. In the subsections that follow, we describe the process of creating the document collection for a given search term and context input pair.

3.1 Building the base document collection

The first step in the document creation process was actually finding the different definitions that we wanted to consider for the input search term. We looked into several websites as potential data sources, and we found that dictionary.com contained a comprehensive list of definitions for most search terms, and was simple enough to parse to be a feasible starting place. It also contained a small number of example sentences for each definition, which would be integral to the construction of the documents.

Therefore, the system starts by parsing the webpage returned after searching the input search term. It maintains a list of the definitions found on the page, and builds the initial document collection where each document contains the text of the definition along with the example sentence(s) that came with it. Due to the well-formatted HTML of the website, the system is able to accurately parse all the different definitions and the example sentences along with them. The range of the number of definitions for a search term is anywhere from 1 to more than 50, with the average being around 10-15.

With the collection of documents indexed by Lucene, the system then uses the Jelinek-Mercer Smoothed Language Model to rank them against the input context. Jelinek-Mercer Smoothing was chosen based on a number of interleave tests among the different ranking models: Okapi BM25, Pivoted Length Normalization, Jelinek-Mercer Smoothing and Dirichlet Prior Smoothing.

At this point, the system functioned in a rudimentary level, functioning for a small set of words with specific contexts that matched the example sentences found in the dictionary website pretty closely. On the contrary, the system suffered with multiple terms which the website either had very short examples or no examples at all. The cause was clearly the limited amount of data from dictionary.com and thus more data was needed.

3.2 Expanding the data

In order to improve the system's performance, we needed to increase the document size by finding more relevant information for each definition. To do so, webpages from thesaurus.com and oxforddictionaries.com were also parsed. The thesaurus.com website contains synonyms for a given term, sectioned by different the different definitions that these synonyms correspond with. The oxforddictionaries.com website contained a large number of example sentences compared to all other websites, which was ideal for our system as it increased the effectiveness of the Language Model scoring method.

With the addition of data sources, however, a new issue arose. Identical definitions for a term could be worded differently by different websites. For example, for a definition of the verb form of the word "bear," one website could contain "Support; Carry the weight of" while another could contain the equivalent definition: "To hold up; Support." We wanted the new data to expand the document size of existing definitions, not add new definitions. As such, we needed a way to map the new definitions onto the old ones. Our solution to this problem was to run the same type of search on the indexed documents with the query being the "new" definition. Once the best-matching "old" definition was found, we expanded that definition's document to contain the new data. This is an imperfect process, and definitely introduces the possibility of error in the system. Regardless, once the documents were expanded, the system indexes them all again. Overall, this expansion-re-indexing process occurs twice, once for each of the two new data sources.

With the data from these new sources added to the documents in the collection, we saw significant improvements in the system's performance. This makes sense as short documents lead to weak and ineffective Language Models. However, we still noticed some significant issues with the system, most importantly that it would sometimes return completely non-relevant definitions of the term in the wrong part of speech. This gave us the inspiration to adapt the system as described in the next subsection.

3.3 Part-of-Speech Tagging

In order to make further improvements to the system, we had to take a different approach from simply gathering more data. From a few test runs, we noticed that oftentimes noun definitions of a word would pop up in our results list even when the word was clearly being used as a verb. The reverse was also true. This led us to realize that we could make use of the context in another way, namely by using it to determine the part of speech of the word. We could then use the part of speech of the word in context to narrow down the list of possible definitions that we were considering, thus increasing the chance of ranking the definitions correctly.

Luckily, we didn't have to implement the part-of-speech identifier ourselves, instead, the Apache OpenNPL Part-of-Speech tagging package was used to implement this improvement. The package, when given a tokenized sentence, returns each token mapped to a part-of-speech tag, along with the probability of correctness. Therefore, to use it, we first tokenized the context into space and punctuation separated tokens, and then fed it to the package. As its predictions aren't 100% accurate, we decided only to use the predicted part-of-speech to filter the definitions to be considered if the probability of correctness was above 0.9. We arrived at this number from a series of empirical observations. In general, the package classifies nouns, verbs, adjectives, and adverbs with upwards of 99% confidence, and since these are the main inputs we expect for our system, we are confident in the effectiveness of adding this component to the system.

3.4 Multiple occurrences of the search term

As we continued to test the system, we came across one final issue that we had to combat: sometimes a word is used more than once in a context with different meanings or parts of speech each time. Up to this point, we'd found the part of speech of the word by searching the context for it. However, with multiple occurrences of the word, we would require the user to tell us exactly which word it was that they were inquiring about. For this system, we just solved this issue by displaying their input back to them and having them choose the number corresponding with the instance of the word they were interested in. Figures 3 and 4 show how the results of the system are completely different depending on which word the user selects, even though the sentences are exactly the same.

```
Enter your Sentence: The brown bear couldn't bear the pressure of the black bear
Enter your search term: bear
Search term occurred multiple times in context, please specify which:
the brown [1.bear] couldn't [2.bear] the pressure of the black [3.bear]
1
1. any of various bearlike animals, such as the koala and the ant bear
2. any plantigrade mammal of the family ursidae: order carnivora (carnivores).
  bears are typically massive omnivorous animals with a large head, a long shaggy
  coat, and strong claws see also black bear, brown bear, polar bear related
  adjective ursine
3. any of various animals resembling the bear, as the ant bear.
```

Figure 3: Here the user selects the first occurrence of the word "bear," where it is used as a noun.

```
Enter your Sentence: The brown bear couldn't bear the pressure of the black bear
Enter your search term: bear
2Search term occurred multiple times in context, please specify which:
the brown [1.bear] couldn't [2.bear] the pressure of the black [3.bear]
2
1. to hold up; support
2. to hold or remain firm under (a load)
3. to press or push against
```

Figure 4: Conversely, here the user selects the second occurrence of the word "bear," where it is used as a verb.

4 EXPERIMENT

Once the system was fully developed, it was time to run some experiments to prove its effectiveness. First, we had to determine what data set we were going to use. To do so, we recalled that the initial motivation for the project was to make it easier to comprehend scientific literature. Therefore, we decided to use the abstracts from 17 different papers on information retrieval that we found on the course website as our contexts, and from each of the abstracts

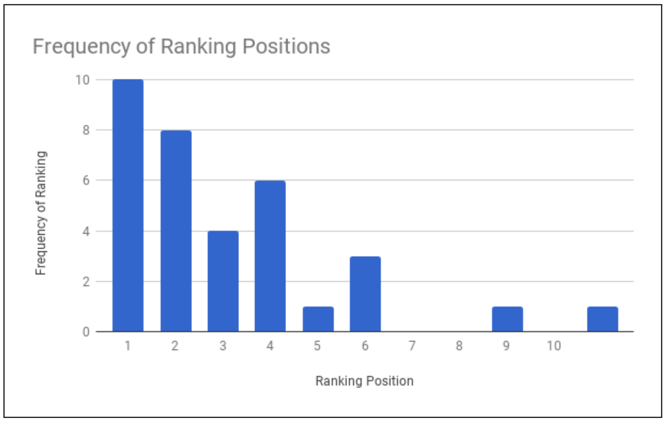

Figure 5: Frequency chart of ranking positions of the correct definition. The one unlabeled bar is for the definition that wasn't found.

we chose 2 different query terms, one common word, preferably a homonym, and one uncommon, technical term. We chose this spread because we wanted to know how the system fares in both scenarios. Overall, there were 34 query term-context pairings. For each query term-context pair, we looked up the term on dictionary.com and identified the definition that we believed corresponded best with the meaning of that word in context. Then, we proceeded to run the system on the query term-context pair and record the reciprocal rank of the correct definition.

In order to evaluate the system as a whole, we looked at the mean reciprocal rank of the results set, calculated with the formula below.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{1}$$

The mean reciprocal rank for this system in this experiment was 0.518. In all 34 tests, only once was the system unable to return the correct definition of the word within the top 10. Figure 5 shows the frequency of each ranking position in our overall results. The complete data table can be found in Appendix A.

Because there aren't currently any other systems like this to compare to, it is difficult to tell if our system's performance is relatively "good" in a statistical sense. However, we can draw some conclusions if we make a few assumptions about the definitions we get from dictionary.com. Since the definitions on dictionary.com aren't ordered in any way relative to context-relevance, we assume that the probability of the correct definition being located at any one ranking position follows a uniform distribution. We can therefore calculate the expected reciprocal rank of a specific trial using just the number of definitions of the query term in that trial. If we let this number be n then we can calculate the expected reciprocal rank with the formula below.

$$ERR = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{i} \tag{2}$$

Using this equation, we were able to determine the expected reciprocal ranks for each trial. The complete data table of observed RRs and ERRs can be found in Appendix A. In looking at the two sets of reciprocal ranks side-by-side, the observed RR was greater than the ERR in 30 out of the 34 trials. Therefore, to test for statistical significance, we performed a one-tailed paired T-test. The p-value resulting from this test was 1.37×10^{-6}, indicating that there is in fact a positive statistically significant difference between the observed RRs and ERRs from this experiment. This allows us to confidently assert that our system is substantially objectively better at providing context-relevant term definition compared to a system that does no consideration of context at all.

5 LIMITATIONS

There are still a number of limitations in our system that undoubtedly impact its ability to perform optimally. First of all, the system is limited by the fact that it will only return definitions listed on dictionary.com. Despite dictionary.com having upwards of 50 definitions for some words, there are certainly others for which there are no definitions at all. This was the case for one of the words in our test data set, "deployment." Interestingly, dictionary.com only contains entries for the word "deploy," which led to no results being returned by our system. In a future iteration of this system, we would need a more comprehensive source of definitions to initialize the system with.

A second obvious limitation that the current implementation of the system has is in its method of growing the document data. Since it does this by using the markedly non-perfect method of repeated searching the index and expanding the document that matches the new data best, there is lots of room for error. It is certainly possible that in some cases, adding data from other sources in this "maximum-likelihood" manner, would hurt the results more than it helps them if the definitions from the different sources are matched incorrectly.

A third limitation in the system is that our ability to use data from a given data source is entirely dependent on our ability to parse the data from the webpage. We considered using wiktionary.com as our original source of definitions, as it seemed a little bit more comprehensive than dictionary.com for some search terms, but we were unable to parse the website in a generic way due to the unorganized source code. Because our documents are generated on the fly, we cannot accept the possibility of the system to fail to parse a webpage for any term. In other words, since we needed to guarantee the system's robustness under all inputs, we needed to gather data from a reliably formatted webpage.

A fourth limitation in the current system is its inability to handle multiple-word search queries. It is conceivable that in the future, we would want to be able to search for a phrase or other set of words in context. In order to extend the current system to support this kind of search, however, we would have to significantly restructure the system and rethink the use of part-of-speech tagging, which proved so helpful in the single-word-query case. A possible a approach to solving this limitation could be to incorporate translation tools such as Google Translate for multiple-word searches, instead of limiting our choices to dictionary websites.

6 FUTURE WORK

Along with addressing the aforementioned limitations, there are still plenty of ways that the system could improve with future work. The first, most obvious way to improve the system would be to add more data sources. Having more data sources leads to having longer documents from which more effective language models can be built. This in turn leads to better rankings lists of definitions, exactly what our system aims to improve.

A second major area for improvement of the current system is in our analysis and usage of the context. Right now, the input context is used for two things–to determine the part of speech of the search term and to construct a language model that we believe contains embedded information about the correct definition of the search term within. However, the current system relies on the user to specify the context, and we treat all words in the context equally. It is plausible, however, that the system would benefit from weighing the words differently based on some metric. For instance, it might make more sense to weigh more heavily words that are in close proximity to the search term. It might also make sense to weigh more heavily terms that are of the same part-of-speech as the search term. These proposals would need to be tested empirically to determine their effectiveness, but it stands to reason that they would improve the system's performance.

A third area where future work could improve the system is the aggregation of definitions. Removal or merging of similar definitions from both different data sources as well as the original data source itself might allow the system to classify between the different and truly distinct definitions easier. For instance, dictionary.com returns, as different definitions of the verb bear, "to hold up; support" and "to hold or remain firm under (a load)" which seem almost identical and interchangeable. It would be beneficial if the system could recognize that these two definitions really represent the same thing, and then merge them and their corresponding documents together. Since we cannot control how many of these similar definitions occur for any given search term, this merging could prove very helpful in eliminating a lot of the clutter. This would definitely improve the system's discriminatory power.

A fourth area where the system might benefit from future work would be to add some user feedback and personalization capabilities. If the user could indicate which definition they liked the best out of the returned results, we could construct a profile for that user and perhaps learn information about them such as what data source they prefer their information to come from.

This last proposal ties in well with the idea that this system could be transformed into a browser extension in the future. As a browser extension, the system would be extremely useful for on-the-fly context-relevant look-ups, and it would be very intuitive to use. The system could copy the same user interaction scheme as the MacBook "look up" function, and doing so would eliminate the need to choose the context explicitly. It would also eliminate the need to select the word within the context if it was used multiple times, as the user would have already selected the word itself directly. In this scenario, the context could be gathered from the HTML of the document as a whole, and then some of the weighting schemes proposed above could be applied.

7 CONCLUSION

In this paper we have presented a novel system for providing context-relevant term definition. We believe the need for such a system is present and rising rapidly, as the world's interconnected population continues to diversify. The system we've developed performs statistically significantly better than random definition ordering in terms of ranking definitions by their contextual relevance. Furthermore, the mean reciprocal rank in our experiments was 0.518, and the mode ranking position of the correct definition was 1, both statistics indicating good system performance. Despite our success, we've still highlighted a number of important drawbacks and limitations that the system has in its current implementation. Fixing some of these limitations as well as adding the modifications discussed in Section 6 could make this system marketable to the public, as we've already proven its usefulness and effectiveness. Public access to this system in its final form would benefit everyone from elementary school students to non-English speaking researchers.

ACKNOWLEDGMENTS

The authors would like to thank Professor Wang for discussing our project idea with us and for teaching us throughout the course how to use and apply all the necessary components to make this system work.

Appendices

A DATA TABLE

Query	Correct Def	Observed RR	Expected RR	Number of definitions
optimum	a condition, degree, amount or compromise that produces the best possible result	0.5	0.46	5
far	at or to a great distance; a long way off; at or to a remote point	0.25	0.21	16
elaborate	worked out with great care and nicety of detail; executed with great minuteness	1	0.31	9
article	a written composition in prose, usually nonfiction, on a specific topic, forming an independent part of a book or other publication, as a newspaper or magazine.	0.16	0.17	22
semantic	of, relating to, or arising from the different meanings of words or other symbols	0.5	0.46	5
classes	any division of persons or things according to rank or grade	0.11	0.09	47
Deployment		0	0	0
Syntagmatic	pertaining to a relationship among linguistic elements that occur sequentially in the chain of speech or writing	1	0.61	3
crawling	also, spider. digital technology. to retrieve (data) from a website using a computer program, as in order to index web pages for a search engine	0.25	0.15	25
glance	at a glance, from one's first look; immediately	0.16	0.2	17
graph	(maths) a drawing depicting a functional relation...	0.5	0.21	16
matrix	(maths) a rectangular array of elements set out in rows and columns	1	0.14	29
queries	asked or inquired about	0.5	0.24	13
robust	strong and effective in all or most situations and conditions	1	0.34	8
schemes	a systematic plan for a course of action	0.5	0.24	13
order	a condition in which each thing is properly disposed with reference to other things and to its purpose; methodical or harmonious arrangement	0.33	0.05	95
major	greater in size, extent, or importance	0.25	0.19	18
adapted	to make suitable to requirements or conditions; adjust or modify fittingly	1	0.52	4
evaluation	an act or instance of evaluating or appraising.	1	0.75	2
authoritative	having due authority; having the sanction or weight of authority	1	0.41	6
morphological	the form and structure of words in a language	1	0.27	11
roots	the source or origin of a thing	0.33	0.09	49
toy	something that serves for or as if for diversion, rather than for serious practical use.	0.2	0.21	16
network	2. any netlike combination of filaments, lines, veins, passages, or the like	0.5	0.19	18
solution	2. a particular instance or method of solving; an explanation or answer	0.5	0.21	16
developed	4. to come or bring into existence; generate or be generated	0.25	0.16	23
postulate	1. to claim or assume the existence or truth of, especially as a basis for reasoning or arguing.	1	0.26	12
rank	6. to arrange in ranks or in regular formation	0.16	0.1	45
Issue	2. a point, the decision of which determines a matter	0.5	0.11	40
Benchmark	3. a standard of excellence, achievement, etc., against which similar things must be measured or judged	0.33	0.29	10
approach	3. the method used or steps taken in setting about a task, problem, etc.	0.33	0.14	27
address	1. to deal with or discuss	1	0.1	43
function	a relation between two sets that associates a unique element	0.25	0.19	19
employ	4. to make use of (an instrument, means, etc.); use; apply	0.25	0.31	9