論文寫作的理論、實務、參考資源

國立清華大學資訊工程系

張俊盛

逢甲大學圖書館三樓資訊素養教室

2018 0507 14:00 - 16:00

大綱

1 論文的 IMRD 結構
- 論文有固定的篇章結構：簡介、方法、結果、討論
- 明確標示編號、標題

2 寫作功能（Writing Function）與論述區塊（Argumentative Zoning）
- 常見的修辭方式，用來組織論文，讓論述有條理、合邏輯、易理解

3 文步分析
- 論文章節（如簡介）內部的隱藏結構
  * 問題（背景、舉例）
  * 解答（方法、舉例、意見）
  * 實驗、結果、應用
  * 組織

4 學術關鍵詞與搭配詞

- 學術論文的用字不同與口語或非正式文章
  - Academic Keyword List
- 詞彙之間有習慣性的配合（搭配）
  - Academic Collocations List

5 常見文法錯誤
- 非母語寫作者容易受母語影響而有固定常見類型錯誤
- 常見錯誤：冠詞、單複數、補語（動、名、形容詞之後的固定句型）
- **have difficulty in doing something**
- not * **have difficulty to do something**

6 寫作參考工具
- WriteAhead
- Linggle

寫作考試如何評分

- 詞彙
  - 詞彙量、詞彙範圍：通用詞表 GSL vs. 學術詞表 AWL, 虛詞 function words vs. 實詞content words
  - 正確使用高品質詞彙（句法、搭配、風格）
  - 慣用、特殊詞彙
  - 使用同義詞、替代詞，以避免重複
  - 詞彙多元性
  - 精確使用詞彙（選詞、型態、拼字、單複數、時態）

- 文法
  - 正確使用多元句子結構（簡單句、複雜、複合句）
  - 正確使用名詞、動詞、形容詞的補語結構
  - 合理使用風格元素（平行結構、反覆强調等、長短、句型變化）

- 文脈
  - 每個段落都有主題句（先廣後專）
  - 客觀平衡、流暢連貫表達思想、內容
  - 使用各種强化連貫性、邏輯性的機制
  - 重述 Paraphrasing
  - 句子段落間有各式轉折、連接詞（但不重覆、生硬）
  - 指涉前後，預告回顧，圖文並茂加强效果

- 內容充實
  - 具備必要的部份（目的、簡介、文獻、方法、結構、討論、 結論）

Source: Hong Kong University of Science and Technology `cle.ust.hk/tests/elpa/elpa-written-test/`

1 論文的IMRD結構

- 標題（與作者）
- 摘要（與關鍵詞）
- 簡介（Introduction）與相關研究
- 方法（Method）
- 結果（Results）
- 討論（Discussion）與結論
- 誌謝辭
- 參考文獻
- 附錄

IMRD結構範例-1 標題、摘要、簡介、相關研究

Learning Search Engine Specific Query Transformations for Question Answering

Eugene Agichtein *
Columbia University
New York, NY 10027
eugene@cs.columbia.edu

Steve Lawrence
NEC Research Institute
Princeton, NJ 08540
lawrence@research.nj.nec.com

Luis Gravano
Columbia University
New York, NY 10027
gravano@cs.columbia.edu

ABSTRACT

We introduce a method for learning query transformations that improves the ability to retrieve answers to questions from an information retrieval system. During the training stage the method involves automatically learning phrase features for classifying questions into different types, automatically generating candidate query transformations from a training set of question/answer pairs, and automatically evaluating the candidate transforms on target information retrieval systems such as real-world general purpose search engines. At run time, questions are transformed into a set of queries, and re-ranking is performed on the documents retrieved. We present a prototype search engine, *Tritus*, that applies the method to web search engines. Blind evaluation on a set of real queries from a web search engine log shows that the method significantly outperforms the underlying web search engines as well as a commercial search engine specializing in question answering.

Keywords

Web search, query expansion, question answering, information retrieval

1. INTRODUCTION

A significant number of natural language questions (e.g., *"What is a hard disk"*) are submitted to search engines on the web every search engine, but rather hardware tutorials or glossary pages with definitions or descriptions of hard disks. A good response might contain an answer such as: *"Hard Disk: One or more rigid magnetic disks rotating about a central axle with associated read/write heads and electronics, used to store data..."*. This definition can be retrieved by transforming the original question into a query {*"hard disk" NEAR "used to"*} . Intuitively, by requiring the phrase *"used to"*, we can bias most search engines towards retrieving this answer as one of the top-ranked documents.

We present a new system, *Tritus*, that automatically learns to transform natural language questions into queries containing terms and phrases expected to appear in documents containing answers to the questions (Section 3). We evaluate *Tritus* on a set of questions chosen randomly from the Excite query logs, and compare the quality of the documents retrieved by *Tritus* with documents retrieved by other state-of-the-art systems (Section 4) in a blind evaluation (Section 5).

2. RELATED WORK

There is a large body of research on Question-Answering, most recently represented in the Text Retrieval Evaluation Conference (TREC) Question-Answering track [22], which involves retrieving a short (50 or 250 byte long) answer to a set of test questions. In our work we consider a more general class of questions, where the answers may not be short, precise facts, and the user might be in-

Source: Agichtein, Eugene, Steve Lawrence, and Luis Gravano. "Learning to find answers to questions on the web." *ACM Transactions on Internet Technology* (*TOIT*) 4.2 (2004): 129-162.

IMRD結構範例-2 方法

3. THE TRITUS SYSTEM

Submitting natural language questions (e.g., *"How do I tie shoelaces?"*) to search engines in their original form often does not work very well. Search engines typically retrieve documents similar to the original queries. Unfortunately, the documents with the best answers may contain only one or two terms from the original queries. Such useful documents may then be ranked low by the search engine, and will never be examined by typical users who do not look beyond the first page of results. To answer a natural language question, a promising approach is to automatically reformulate the question into a query that contains terms and phrases that are expected to appear in documents containing answers to the original question.

3.1 Problem Statement

We focus on the first step of the question answering process: retrieving a set of documents likely to contain an answer to a given question. These documents are then returned as the output of the system. The returned documents can be examined by a human user directly, or passed on to sophisticated answer extraction modules of a question answering system (e.g., Abney et al. [2000], Mann [2002], Prager et al. [2002], and Radev et al. [2002]). Thus, it is crucial that the

IMRD結構範例-3 實驗、評估、結果

4. EXPERIMENTAL SETTING

Tritus was designed to retrieve documents from the Web that are likely to contain answers to a given natural language question. As such, *Tritus* will be trained and evaluated over the *Web at large*. Furthermore, since the goal of *Tritus* is to retrieve a good set of *documents* (as opposed to extracting an exact answer), we evaluate *Tritus* on the document level. Finally, since we do not restrict the type of questions that users can ask, we will use real human judges to evaluate the quality of documents retrieved by *Tritus*. In this section, we first present the details of training *Tritus* for the evaluation (Section 4.1). Then, Section 4.2 lists the retrieval systems that we use in our comparison. Section 4.3 introduces the evaluation metrics for the performance of the retrieval systems, and details of the queries evaluated and relevance judgments are reported in Section 4.4.

4.1 Training Tritus

We used a collection of approximately 30,000 question-answer pairs for training, obtained from more than 270 Frequently Asked Question (FAQ) files on various subjects. Figure 4 shows a sample of the question-answer pairs. We obtained these FAQ files from the FAQFinder project [Burke et al. 1995]. All of the FAQ files used for evaluation are publicly available in parsed form.[5] We evaluated four question types. The number of question-answer training pairs in the collection for each of the question types is shown in Table VI.

IMRD結構範例-4 評估、結果

5. EVALUATION RESULTS

In this section, we report the results of the experimental evaluation using the methodology described in the previous section. First, in Section 5.1 we report the results of our original evaluation performed in the Fall of 2000, which totalled 89 questions evaluated by volunteers, mostly acquaintances and colleagues of the authors who were requested to help with the evaluation. In Section 5.2 we present the results of the new, more extensive evaluation performed in the Spring of 2002, which additionally involved anonymous and unknown judges that participated in evaluating all of the updated systems as described above.

5.1 Results from the Original (2000) Evaluation

During this evaluation, 89 questions were evaluated by volunteer judges. Table X lists the number of questions of each type that were presented to the judges and the number of questions that were actually evaluated by the judges.

Figure 7(a) shows the average precision at K for varying K of **AJ**, **AV**, **GO**, **TR-GO**, and **TR-AV** over the 89 test questions. (**TR-ALL** is new and was not part of the 2000 evaluation.) As we can see, *Tritus*, optimized for Google, has the highest precision at all values of document cutoff K. Also note that both

IMRD結構範例-5 討論、結論

6. FUTURE WORK AND SUMMARY

Many avenues exist for future research and improvement of our system. For example, existing methods for extracting the best passages from documents could be implemented. Domain knowledge, heuristics, and natural language parsing techniques could be used to improve the identification of question types. Multiple transformations could be combined into a single query. Questions could be routed to search engines that perform best for the given question type. Additionally, an interesting direction to explore is creating phrase transforms that contain content words from the questions. Yet another direction of research would be to make the transformation process dynamic. For example, transformations where we expect high precision could be submitted first. Based on the responses received, the system could try lower precision transforms or fall back to the original query.

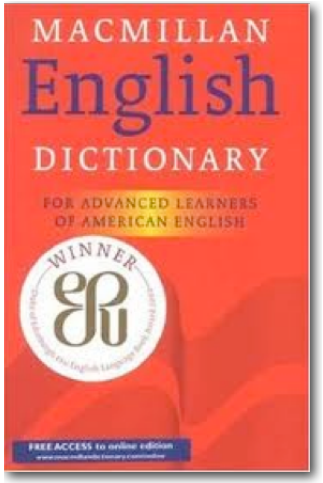
In summary, we have introduced a method for learning query transformations that improves the ability to retrieve documents with answers to questions using an information retrieval system. The method involves classifying

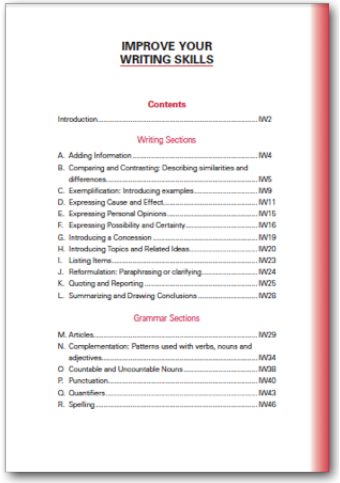
如何安排章節內部的段落句子

- 如何安排章節內部的段落句子
  - 先廣泛後特定（From General to Specific）
  - 先承接後引介（From Old to New)
  - 先主題後內容（Topic Sentence in a paragraph）
  - 先時空後人物、事件（Locating time and space）
  - 修辭功能（內容、順序）（將在 第 2 部分 介紹修辭功能
    * 添加强化、舉例説明、比較對照
  - 論文特有的修辭功能（將在 第 3 部分 介紹文步）
    * 有如章節結構：背景（簡介）、缺口（問題）、方法（解答）等

2 寫作功能與修辭類型

- Macmillan English Dictionary for Advanced Learner (MED) comes with
  - **Improve Your Writing Skills** (Granger et al. 2007)
    - 12 Writing Sections + 6 Grammar Sections (Common Errors)




IMPROVE YOUR WRITING SKILLS

Contents

Introduction IW2

Writing Sections

A. Adding Information IW4
B. Comparing and Contrasting: Describing similarities and differences IW5
C. Exemplification: Introducing examples IW9
D. Expressing Cause and Effect IW11
E. Expressing Personal Opinions IW15
F. Expressing Possibility and Certainty IW16
G. Introducing a Concession IW19
H. Introducing Topics and Related Ideas IW20
I. Listing Items IW23
J. Reformulation: Paraphrasing or clarifying IW24
K. Quoting and Reporting IW25
L. Summarizing and Drawing Conclusions IW28

Grammar Sections

M. Articles IW29
N. Complementation: Patterns used with verbs, nouns and adjectives IW34
O. Countable and Uncountable Nouns IW38
P. Punctuation IW40
Q. Quantifiers IW43
R. Spelling IW46

References

1. De Cock, S., Gilquin, G., Granger, S., Lefer, MA., Paquot, M., and Ricketts, S. (2007) *Improve your writing skills.* In M. Rundell (editor in chief) *Macmillan English Dictionary for Advanced Learners* (second edition) IW1–IW50.
2. UNC Writing Center Handouts: Transitions`writingcenter.unc.edu/tips-and-tools/transitions/`

Writing Sections: Rhetoric Functions

Function	Code	Example
1. Add information	add	*Additionally,*
2. Describe similarities/differences	CTR	*Similarly,*
3. Introduce an examples	exam	*For example,*
4. Express cause and effect	cause (effect)	*As a result,*
5. Voice opinions or evaluate	opin (eval)	*Intuitively,*
6. Express possibility/certainty	poss	*may, must, could*
7. Hedge or concede	hedge	*however, although*
8. Introduce topics	AIM	*In this paper, we ...*
9. List Items	list	*First, ... Finally, ...*
10. Paraphrasing or clarifying	para	*In other words,*
11. Cite others	OTH (report)	*X presents ...*
12. Conclude and summarize	sum	*In summary,*

Writing Sections: Rhetoric Functions

Function	Code	Example
13. Pointing out a problem	problem	*Unfortunately,*
14. Propose a solution	solution	*A promising approach is*
15. Emphrasize	emph	*... automatically ... automatically*
16. Locate in space	space	*In a research area closer to ...*
17. Locate in time	time	*Recently, after, once, then*
18. Input (AIM)	input	*We are given ..., is the input*
19. Output (AIM)	output	*is returned, output, The goal is*
20. Talk about numbers	amount	*increase by a factor of two*

Writing Sections (1-3)

Function	Adverb/Conj.	Det./Adj.	Prep.	Verb	Noun	Avoid
add	additionally, moreover, furthermore	another	In addition to,	—	—	and, besides
CTR (sim)	similarly, likewise, in the same way	analogous, common, comparable, identical, parallel, similar	like	resemble, correspond	resemblance, similarity, parallel, analogy	look like, like
CTR (diff)	in contrast, in comparison; while, whereas	contrasting, different, differing	in contrast to, in comparison with, unlike	contrast, differ	contrast, difference, distinction	—
exam	for example, for instance, e.g., notably	—	such as, like	illustrate, exemplify	example	—

Writing Functions (4-6)

Function	Adv/Conj/Aux	Det./Adj.	Prep.	Verb	Noun	Avoid
cause (or effect)	Therefore, Thus; so that	because, since	because of, due to; as a result of, as a conseq. of	CAUSE, FOLLOW	cause, factor; effect, result	—
opin (or eval)	In my opinion, In my view, OPIN-ADV	OPIN-PAT	—	seem, prove	opinion, view	—
poss	may, must; probably, obviously	possible, certain	—	—	assumption, belief	—

CAUSE=[V n] allow us to v, bring about, contribute to, generate, give rise to, lead to, result in, yield

FOLLOW=arise from, derive, emerge, follow, result, stem

OPIN-PAT=it is Adj. to, it is Adj. that, it is worth doing,

OPIN-ADV=Interesting, Significantly, Unfortunately, Intuitively, Surprisingly

Writing sections (7-9)

Function	Adverb/Conj.	Det./Adj.	Prep.	Verb	Noun	Avoid
hedge	however, nevertheless, nonetheless; yet; (surp.) although, though; albeit		despite, in spite of, notwithstanding	—	—	—
AIM	incidentally	another; further, last, next; etc, and so on	—	consider, discuss, examine	TOPIC goal purpose aim subject issue	—
list	first, firstly; second, third, fourth, finally	first, second, third, fourth, final, last	—	—	phase, stage, step	firstly, last but not least

Writing sections (9-12)

Function	Adverb/Conj.	Verb	Noun	Avoid
para	in other words; namely, viz., more precisely/accurately, or rather (correction)	i.e., that is, that is to say	cause, factor	viz.
OTH (or report)	correctly/rightly REPORT	REPORT	conclusion, belief; view, opinion (PAT)	in X's view/opinion
sum	In summary, To summarize; In conclusion,	—	—	In sum, Summing up, To sum up, To conclude

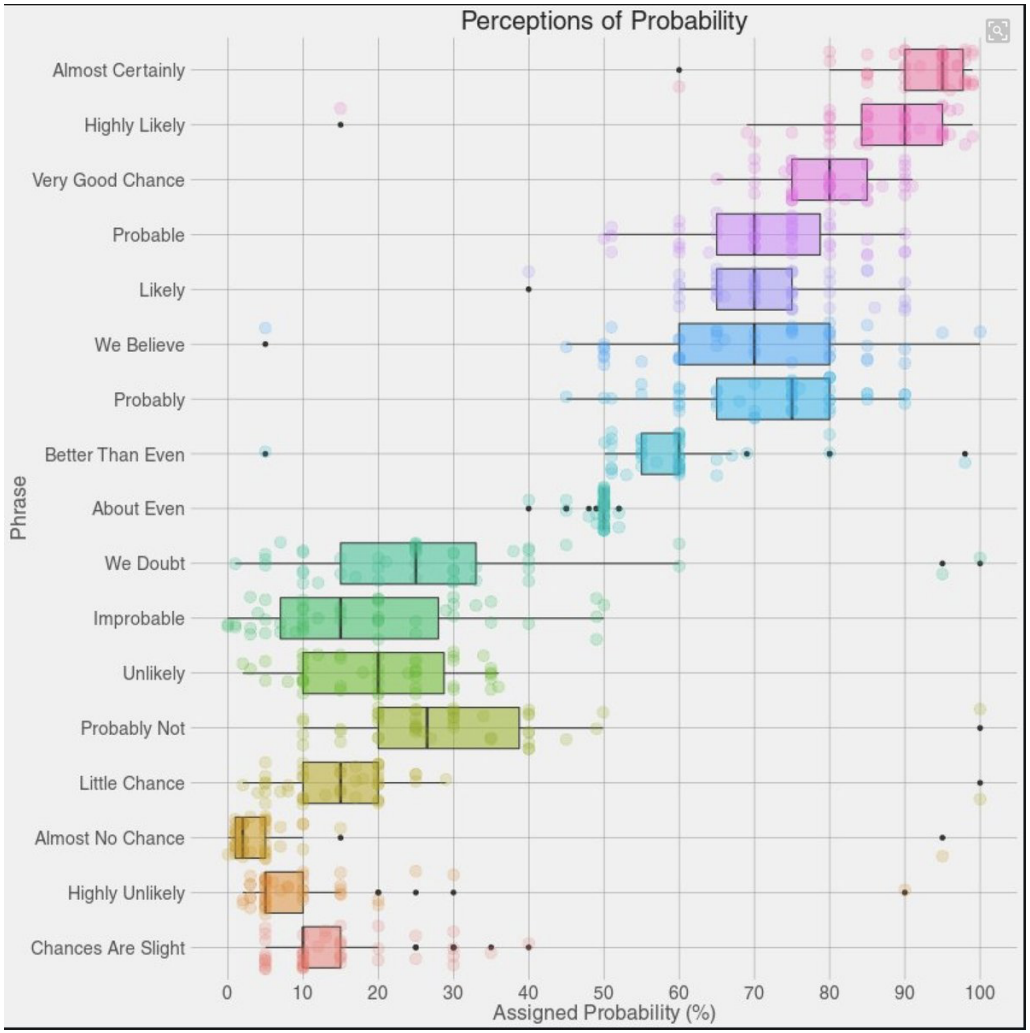
Simple REP: comment, conclude, remark, report, say, write
Interpret: acknowledge, admit, argue, assert, claim, maintain, recognize, stress
REP with findings: analyse, compare, describe, discuss, explain, focus on, show
Agree with REP: as REP, as REPORTed by AUTHOR
Others: add, assume, believe, comment, concede, conclude, confirm, dispute, emphasize, estimate, find, indicate, mention, note, observe, point out, propose, state, suggest

PAT = X's view/opinion, according to X, X hold the view that, X is of the opinion that

Writing sections (13-19)

Function	Adverb/Conj.	Verb	Noun	Avoid
problem	in other words; namely, viz., more precisely/accurately, or rather (correction)	i.e., that is, that is to say	cause, factor	viz.
solution	correctly/rightly REPORT	REPORT	conclusion, belief; view, opinion (PAT)	in X's view/opinion
emph	In summary, To summarize; In conclusion,	—	—	In sum, Summing up, To sum up, To conclude
space	In summary, To summarize; In conclusion,	—	—	In sum, Summing up, To sum up, To conclude
time	In summary, To summarize; In conclusion,	—	—	In sum, Summing up, To sum up, To conclude

How to hedges



Almost certain
Highly likely
Very good chance
Probable
Likely
We believe
Probable
Better than even
About even
We doubt
Improbable
Unlikely
Little chance
Almost no chance
Almost no chance
Highly unlikely
Chances are slim

This is how CIA hedges

References

1. pbs.twimg.com/media/Dacxb4TX4AEYwWw.jpg

2. t.co/JqkZoOQeSr

3. www.cia.gov/library/center-for-the-study-of-intelligence/csi-publicat
   books-and-monographs/

4. psychology-of-intelligence-analysis/art4.html

3 論文的修辭分析：文步

- 論文寫作需要特殊的修辭分析
- 摘要的分析：背景、目的、方法、結果、討論
- 簡介的分析：Create a Research Space (CARS)
- 史維爾（John Swales）指出
  - 先描述研究的背景和論文的利基
  - 如同生態環境中的生物尋找生存空間般
  - 論文必須透過特定的修辭結構，來爭取審查委員、讀者的接受與認同。
  - 大部分論文〈簡介〉的三部曲：三文步

簡介的 CARS 分析

1. 描繪研究領域的空間（背景）	（修辭方式）
• 指出課題重要性、中心性	强調重要性
• 介紹相關研究、應用	定義術語
2. 建立研究的利基（問題、缺口）	
• 指出前人不足、缺陷、侷限	凸顯問題
• 延伸前人研究	表達意見、舉例說明
3. 佔據利基（目的、結果）	
• 指出研究目的/條列議題	提出目的
• 宣佈研究結果/價值	提供解答
• 介紹論文的組織（非必要）	概述組織

論文文步 Moves

- 什麼是「文步」：
  - 分段、分析論文的一種架構
  - 就像章節的標題
  - 文步就像隱形的標題，把沒有標題的段落、句子，切割開來、標示其資訊與修辭方式
  - 一般常用的文步：背景、目的、方法、結果、討論（結論）
- Simone Tuffel (1990) 提出的 Argumentative Zoning 的文步架構
  - BKG（黃色）：背景
  - OTH（橘色）：引用前人文獻
  - OWN（藍色）：本研究之方法、結果
  - AIM（粉紅）：本研究之目的
  - TXT (紅色）：參照文字、圖表
  - CTR（綠色）：對照、比較之討論
  - BAS (紫色）：研究間承續關係之討論
- Source: `http://www.cl.cam.ac.uk/~sht25/az.html`

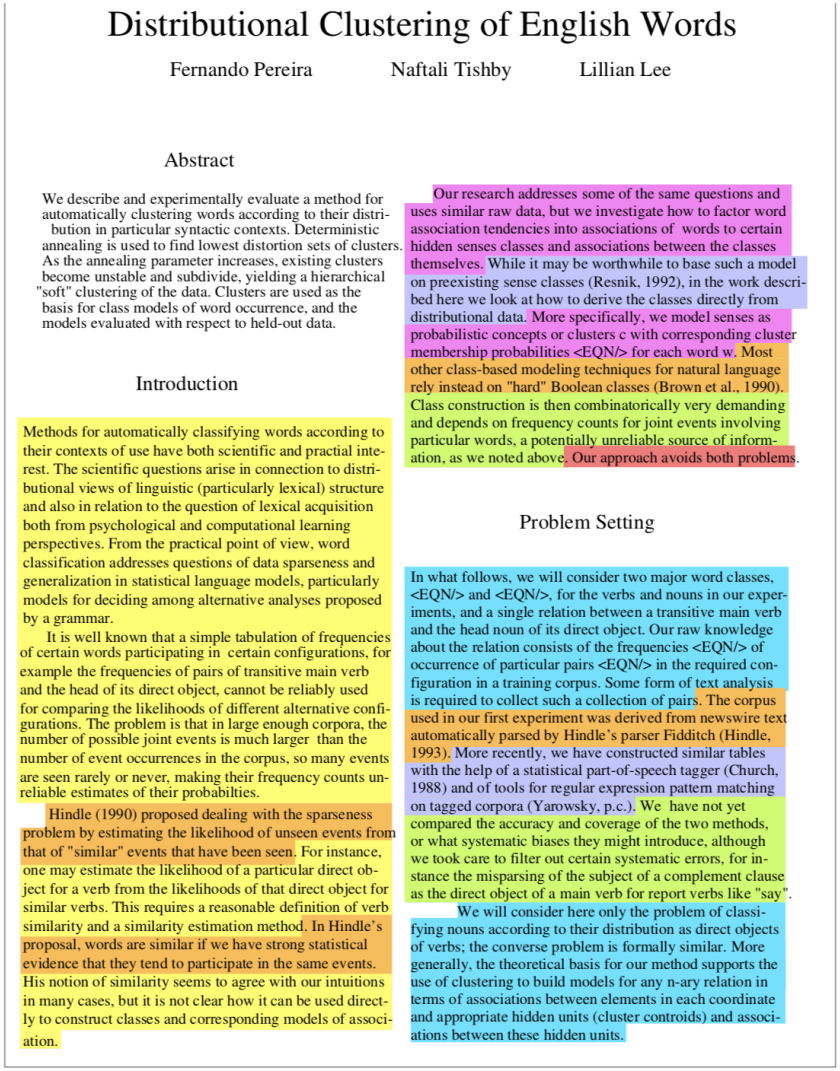
Moves in Papers

BKG (yellow)

OTH (orange)

BKG (yellow)

OTH (orange)



Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

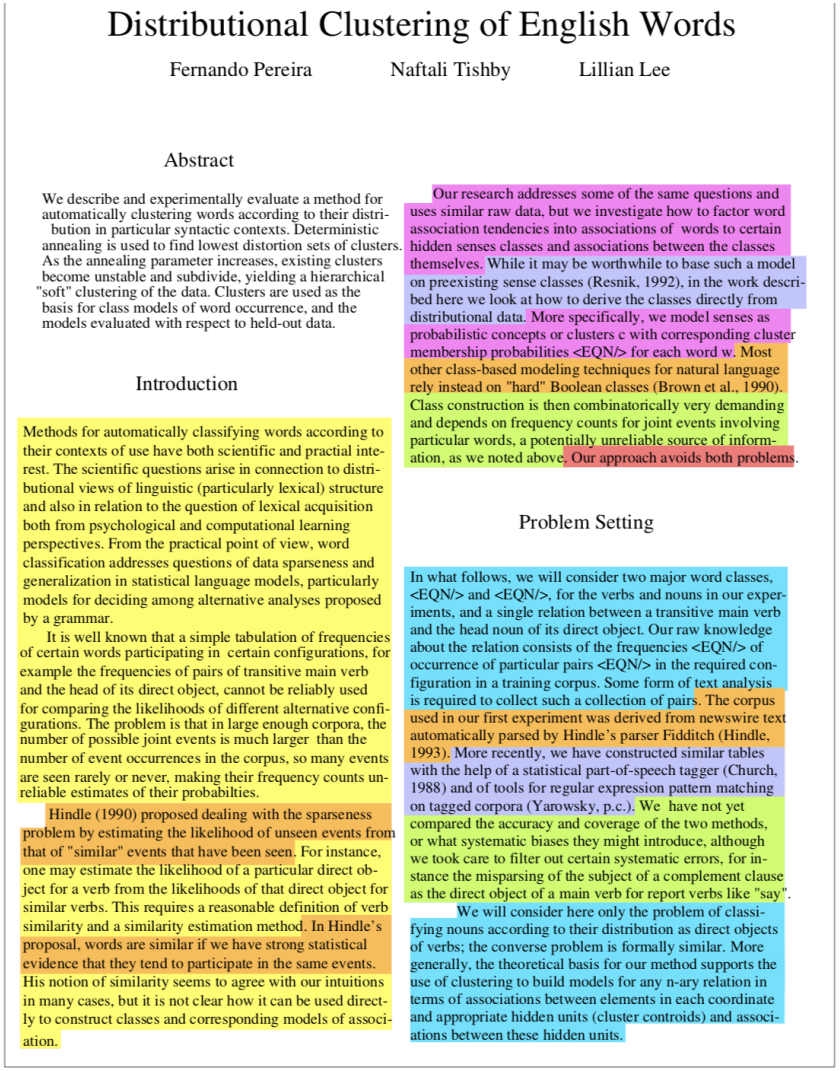
Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster centroids) and associations between these hidden units.

Moves in Papers



Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster controids) and associations between these hidden units.

```
AIM (pink)

BAS (violet)
AIM (pink)
OTH (orange)
CTR (green)
TXT (red)



OWN (blue)
OTH (orange)

BAS (violet)

CTR (green)

OWN (blue)
```

ALG 2004–簡介 (1,2)

BKG [1] ***Many*** natural language questions ***(e.g., "What is a hard disk?")*** are
emph submitted to search engines on the Web ***every day***, and an ***increasing***
exam ***number of*** search services on the Web specifically target natural language questions.

BKG [2] **For example,** AskJeeves (www.ask.com) uses databases of precompiled
exam information, metasearching, and other proprietary methods, ***while***
ctr services***such as*** AskMe (www.askme.com) and Google Answers (answers.google.com) facilitate interaction with human experts.

BKG [1] Web search engines ***such as*** AltaVista (www.altavista.com) and Google
exm (www.google.com) **typically** treat natural language questions as lists of terms and retrieve documents similar to the original query.

gap [2] ***However***, documents with the best answers ***may*** contain ***few*** of the
poss terms from the original query and may be ranked low by the search engine.

poss [3] These queries ***could*** be answered more precisely ***if*** a search engine recognized them as questions.

ALG 2004–簡介 (3)

OWNm ^{1} ***Consider*** the question ***"What is a hard disk?"***.
exam

OWNm ^{2} The best documents for this query are ***probably not*** the company Web
poss sites of disk storage manufacturers, which ***may*** be returned by a general-
para purpose search engine, ***but rather*** hardware tutorials or glossary pages with definitions or descriptions of hard disks.

OWNm ^{3} A good response ***might*** contain an answer ***such as***: "Hard Disk: One or
poss more rigid magnetic disks rotating about a central axle with associated
exam read/write heads and electronics, used to store data. . . ".

OWNm ^{4} This definition can be retrieved by transforming the original question into a query hard disk NEAR "used to" .

OWNm ^{5} ***Intuitively,*** by requiring the phrase "used to", we can bias search engines
opin towards retrieving this answer as one of the top-ranked documents.

ALG 2004–簡介 (4)

OWNr [1] **We present a new system,** Tritus, that automatically learns to transform natural language questions into queries expected to retrieve answers to the
exam question using a given search engine ***(e.g., a specific Web search engine such as Google)***.

TXT [2] **An example Tritus search** for the question "what is a hard disk?" is shown
exam **in Figure 1**.

OWNr [3] Tritus has determined the best 15 transforms for the "what is a" type
exam of question ***(e.g., {hard disk "is usually"}, {hard disk called})*** for the specific underlying search engine (in this case, for Google).

OWNm [4] Tritus learns these effective transformations automatically during training by analyzing a collection of question-answer pairs, and recognizing the
exam indicative answer phrases for each question type ***(e.g., Tritus learns that a phrase "is usually" is a good transform for "what is a" questions)***.

TXT [5] We describe the Tritus training process in more detail **in Section 3**.

ALG 2004–簡介 (5)

OWNm ^{1} *list exam* At run-time, Tritus ***starts with*** a natural language question submitted by the user ***(e.g., "what is a hard disk?")***, which is transformed into a set of new, effective queries for the search engine of interest.

list^{2} Tritus ***then*** retrieves and reranks the documents returned by the underlying search engine.

OWNc ^{3} TXT *ctr* **In our prototype,** Tritus **returns the documents to the user directly** (see Figure 1)**;** ***alternatively,*** the documents returned by Tritus **can be used as input to** a traditional question answering system in order to extract the actual answers from the retrieved documents.

ALG 2004–簡介 (6)

TXT [1] The rest of the article is organized **as follows**.

TXT [2] We review the related work **in the next section**.

TXT [3] **Then** we present our method for automatically learning to transform natural
list language questions into queries containing terms and phrases expected to appear in documents containing answers to the questions (**Section 3**).

TXT [4] As part of our evaluation, we compare the quality of the documents retrieved active researchby Tritus with documents retrieved by other state-of-the-art systems (**Section 4**) in a blind evaluation (**Section 5**) over a set of questions chosen randomly from the query logs of a public Web search engine.

運用寫作樣板

- 閱讀、分析 ALG 2004 的摘要、簡介中的關鍵資訊
- 分析你的研究的對應的關鍵資訊：
  - 領域、重要性（area, importance）
  - 輸入、輸入範例（input, example input）
  - 輸出、輸出範例（input, example input）
  - 困難（problem, gap）
  - 中間結果、路線（intermediates, approach, solution）
  - 方法、步驟、結果（method, steps, results）
  - 章節組織（organization）

寫作前—分析、規劃

重點資訊	範例論文	你的論文
1. importance	many questions	
2. example systems	Google, AskMe	
3. area	Question Answering	
4. input	question	
5. example input	What is a harddisk?	
6. example output	A harddisk is a device	
7. problem	question is not query	
8. solution	question $\rightarrow$ transf.	
9. intermediates	what is $\rightarrow$ used to	
10. system name	Tritus	
11. Step 1	question $\rightarrow$ q-phrase	
12. Step 2	q-ph $\rightarrow$ transformation	
13. Step 3	evaluate transformation	

ALG 2004–簡介 (1,2)

BKG [1] Many _____ <input> (e.g., _____ <example input> are _____ <submitted> _____ <where, situation> every day, and an increasing number of _____ <systems> target _____ <input> (OR provide _____ <service>). 用數量（輸入、網路系統）、時間，來强調重要性

BKG [2] For example, _____ <example system> _____ <works in some way>, while _____ <system> such as _____ and _____ <2 example systems> <work in some way>. 呼應 an increasing number of ，舉出 3 個性質不同的（對照）範例系統

ctr

BKG [1] _____ <systems> such as _____ and _____ <two example systems> _____ <work in some way>. 繼續舉出 2 個範例系統

problem [2] However, _____ <some input/output could present a problem>.
指出範例系統處理輸入，產生輸出時，可能引發的問題

solution [3] _____ <input or output> could be -ed _____ <processed> more _____ <successfully> if a _____ <system> _____ <does something as an approach (generally)>.
指出可行的研究路線（如果系統採取某種作法——僅概述）

ALG 2004–簡介 (3)

OWNm [1] Consider the _____ <input> "_____" <example input>.
舉出一個輸入的實例

OWNm [2] The best _____ <output> for this _____ <input> are probably not _____ <example output>
舉例說明有問題的「輸出」

OWNm [3] A good response might contain _____ <output> such as: _____ <output example>".
舉例說明理想的「輸出」

OWNm [4] This _____ <output> can be _____ <obtained> by -ing _____ <approach>.
說明如何做（研究路線）就可得到理想的「輸出」

OWNm [5] Intuitively, by -ing _____ <approach>, we can _____ <do better>.
暗示（意見）研究路線直覺、容易理解（評價）—如果怎麼做（研究路線）就可得理想「輸出」

ALG 2004–簡介 (4)

OWNr [1] We present a new system, _____ <name>, that _____ <does something as an approach>, expected to _____ <perform better>.
我們呈現一個新的系統，用（研究路線）得到理想的「輸出」

TXT [2] An example _____ <name> _____ <operation> for the _____ <input>, _____ <input example> is shown in Figure __ .
對某輸入的一個範例的操作過程，見於第幾圖

OWNr [3] _____ <name> has _____ the best _____ < generated intermediates> for _____ <input> (e.g., _____ <example input>).
（圖顯示）系統決定了範例輸入的最佳中間結果

OWNm [4] _____ <name> learns these effective _____ <intermediates> during training by -ing _____ <analyzing trainning set> (e.g., _____ <name> learns _____ <example intermediates> for _____ <example input>).
系統以某某方法，學到（如何產生）這些中間結果（例如，哪些中間結果，就對應到哪一類的輸入）

TXT [5] We describe the _____ <name> training process in more detail in Section ____ .
我們會在某一節，說明這個學習的過程

ALG 2004–簡介 (5)

OWNm [1] At run-time, _____ <name> starts with _____ <input> submitted by the user (e.g., _____ <example input>), which is _____ <processed> into _____ <intermediates>.
在執行的時候，系統開始先取得輸入，將之轉換成中間結果

list[2] _____ <name> then _____ <does something> _____ <produce output>.
系統接著就如何如何做，並產生結果

OWNc [3] In our prototype, _____ <name>s returns the _____ <output> to the user directly (see Figure ___); alternatively, the _____ <output> returned by _____ <name> can be used as input to _____ <downstream systems> in order to _____ <do something>.
在我們的雛形，系統有兩種應用：直接回傳給使用者；替代方案則是饋入某系統，繼續處理

ALG 2004–簡介 (6)

,

TXT [1] The rest of the article is organized as follows.
概述本章之外的組織

TXT [2] We review the related work in the next section.
下一章（第二章）我們回顧相關研究

TXT [3] Then we present our method for -ing _____ <processing input> into _____ <intermediates> expected to _____ <condition> (Section 3).
然後，我們提出處理輸入，產生中間結果，預期有助於得到理想輸出（第三章）

TXT [4] In our evaluation, we compare _____ <our system> (Section 4) with _____ <base lines> (Section 5) over a set of _____ <test cases>.
在評估方面，我們用一組測試資料，比較本系統和基線系統的結果（第４章、第５章）

4 學術關鍵詞與搭配詞

- 學術論文的用字不同與口語或非正式文章
  - Academic Keyword List (`http://www.uclouvain.be/en-372126.html`)
- 詞彙之間有習慣性的配合（搭配）
  - Academic Collocations List: 2,469 most frequent and pedagogically relevant lexical collocations in written academic English in Pearson International Corpus of Academic English (PICAE) with some 25 million words.
    * `pearsonpte.com/organizations/researchers/academic-collocation-list`
    * `pearsonpte.com/wp-content/uploads/2014/07/AcademicCollocationList.xls`
  - `www.ozdic.com/collocation-dictionary`

Academic Keyword List

- 355 nouns ability, absence, account, achievement, act, action, activity, addition, adoption, adult, advance, advantage, advice, age, aim, alternative, amount, analogy, analysis, application, approach, argument, aspect, assertion, assessment, assistance, association, assumption, attempt, attention, attitude, author, awareness, ...

- 233 verbs accept, account (for), achieve, acquire, act, adapt, adopt, advance, advocate, affect, aid, aim, allocate, allow, alter, analyse, appear, apply, argue, arise, assert, assess, assign, associate, assist, assume, attain, attempt, attend, attribute, avoid, ...

- 180 adjectives absolute, abstract, acceptable, accessible, active, actual, acute, additional, adequate, alternative, apparent, applicable, appropriate, arbitrary, available, average, basic, ...

- 87 adverbs above, accordingly, accurately, adequately, also, approximately, at best, basically, ...

- 75 others according to, although, an, as, as opposed to, as to, as well as, because, because of, between, both, by, ...

多用學術單字，而且也要合文法、避免錯誤

- accept 的錯誤句型
  - * The company will not **accept to buy** new machines.
  - v The company will not **agree to buy** new machines.
  - * We can't **accept a motorway to be** built through our town. "
  - v We can't **allow a motorway to be** built through our town.

- accept 的文法規則（grammar patterns）與例句
  - **V n**: **accept risks** to cure chronic disease
  - **V n as n**: **accept it as a member** in both A and B
  - **V that**: **accept that** science is in some sense never value free
  - **V n as adj**: **accept Z as true**; **accept language as transparent**

多用學術單字，並符合習慣性的配合

- learn 的錯誤搭配：* learn knowledge → acquire/gain knowledge
- 詞彙之間有習慣性的配合（搭配）
  - Academic Collocations List
    pearsonpte.com/wp-content/uploads/2014/07/AcademicCollocationList.pdf

32	acquire v knowledge n (* learn knowledge)
33	active adj involvement n
34	active adj participant n
35	active adj participation n
36	active adj role n
37 (be)	actively adv involved vpp
38	acutely adv aware adj
... ...	
2468	written adj statement n
2469	younger adj generation n (the last collocation)

製作學術詞文法、搭配小冊——學習、運用 AKL

ability

ability to do something: Tiredness can seriously impair your ability to drive.

above

above all else: Above all else, the government must keep the promises it has made.

absence

in someone's absence : Mark will be in charge in my absence.

absence of: a complete absence of humor

in the absence of something: In the absence of any contrary agreement, the firm accepts full liability.

abstract

abstract idea/concept/principle/notion: Mathematics is concerned with understanding abstract concepts.

accept

accept that: Most scientists accept that climate change is linked to pollution.

generally/widely accepted : His views on genetics are not now widely accepted.

accept blame/responsibility/liability: We cannot accept liability for items stolen from your car.

accept that: For a long time, he simply could not accept that she was dead.

accept someone as something: Mexico was accepted as a member of the OECD in 1994.

accept someone into something: She was desperate for the children to accept her into the family.

自動產生學術詞彙測驗——學習、運用 AKL

8. I have assumed the basic _____ is identical for each image. *

- exposure
- discovery
- resistance
- existence

9. Ingredients in commercially formulated diets _____ from company to company. *

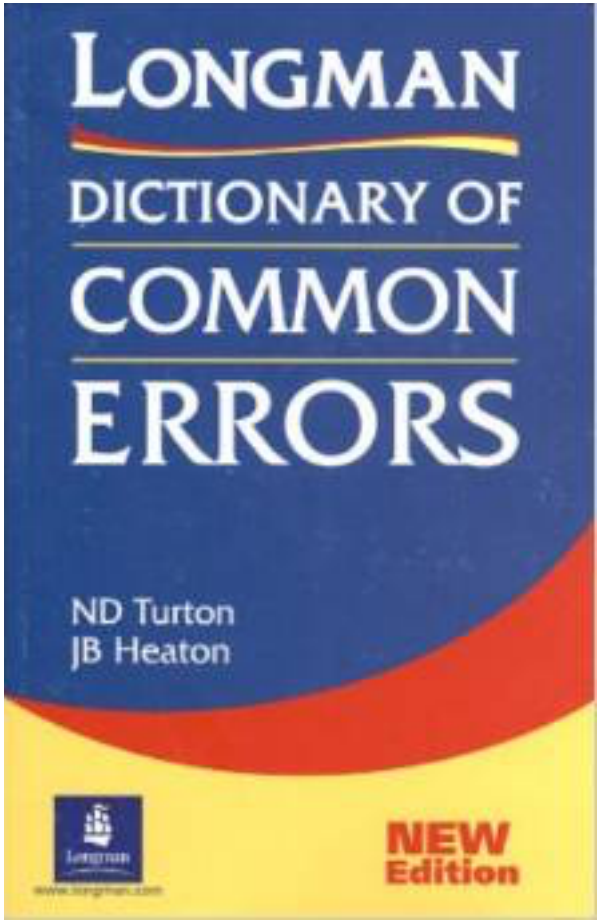
- evaluate
- illustrate
- vary
- concentrate

5 避免、更正文法錯誤

- Grammar Sections (in Improve Your Writing Skills)
  - Articles
  - Complementation: Patterns used with verbs, nouns and adjectives
  - Countable and Uncountable Nouns
  - Punctuation
  - Quantifiers
  - Spelling

常見文法錯誤

- Turton and Heaton (1996)
- 根據學生寫作語料庫的常見錯誤
- 2,500 個易引發錯誤的單字
- 照字母順序排列
- 如 ability: V of -ing → V to-infinitive
- 真實寫作錯誤例句及更正
- 解釋「錯與對」的文法原理



- a, an
  - I hope you all have {*a|an} enjoyable stay.
  - My husband is doing {*a|an} MSc in civil engineering.
  - Sometimes it is difficult to live {*a|an} honest life.
  - The child had been {*a|} deaf since birth.
  - One of the girls I share with is{*a|} British.

- abandon
  - Since capital punishment was *abandoned—abolished, the crime rate has increased.
  - It is difficult to reach *abandoned—remote places such as small country villages.

- ability
  - These machines are destroying our {*ability of thinking|ability to think}.
  - I want to improve my {*ability of reading|reading ability}.
  - I want to improve my ability {*of|in} English.

- able
  - One man is {*able to destroy|capable of destroying} the whole world.

常見文法錯誤 (accept)

1

× The company will not accept to buy new machines.

✓ **The company will not agree to buy new machines.**

*You **accept** someone's advice, opinion, or suggestion BUT you **agree** (= say you are willing) to do something. Compare: 'I accepted her suggestion and agreed to see the doctor that evening.'*

2

× The driver did not accept me to get on the bus.

✓ **The driver did not allow me to get on the bus.**

× We can't accept a motorway to be built through our town.

✓ **We can't allow a motorway to be built through our town.**

*You **allow/permit** someone to do something, or **let** them do it: 'Many parents do not allow/permit*

用什麼形式的文法可以解釋錯誤

- 一般辭典都針對單字，說明相關文法訊息（可屬性、常複數、及物、不及物，搭配介詞）
- 需要一種詞彙化文法 lexical grammar
- Pattern Grammar 是最簡明經濟的詞彙化文法（教學文法）
- PG 提供一個模式來描述單字的句法環境

- 每個單字（實詞）有一組文法規則（patterns）描述單字的用法

- （通常）一個規則，一個語意

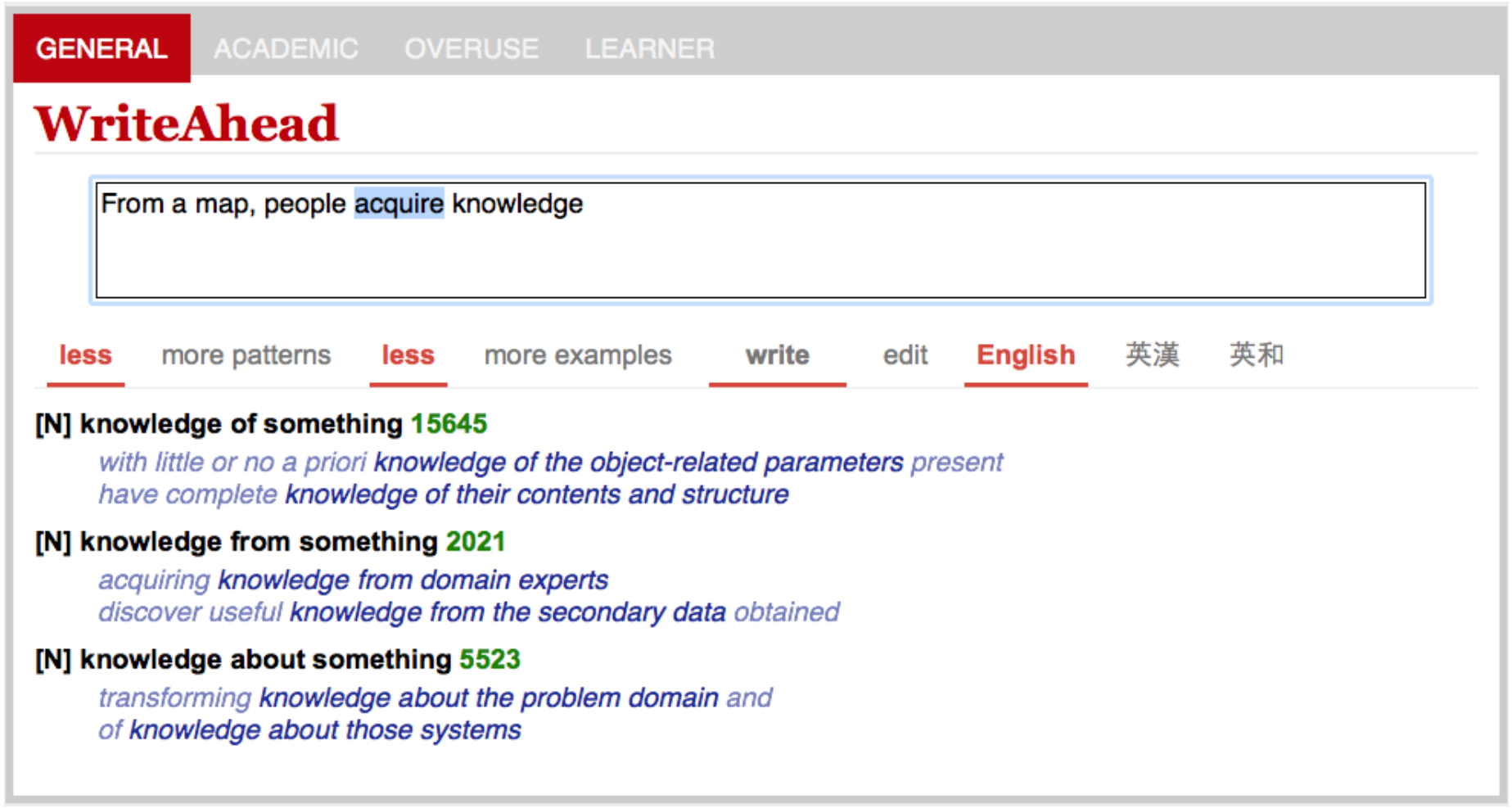
- Sources

  - `http://en.wikipedia.org/wiki/Pattern_grammar`
  - Hunston and Francis (2000): A corpus-driven approach to the *lexical grammar* of English

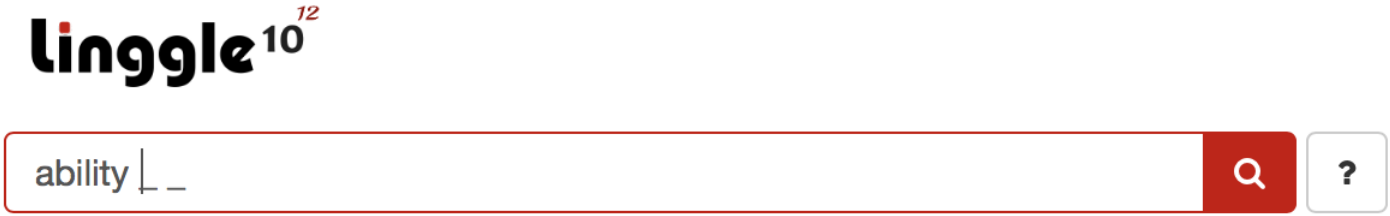
Example of Pattern Grammar

- Skim (v.) includes the following patterns in the COBUILD dictionary
  - **V n off/from n**: ***Skim*** *the fat* ***off*** *the soup.* (limited prep. allow)
  - **V n**: ***Skim*** *the wall surface smooth ready for painting*
  - **V over/across**: *Water skiers* ***skimmed across*** *the bay.*
  - **V through n**: ***Skim through*** *the report and check for spelling mistakes?*

6. WriteAhead Showing Patterns of *knowledge*

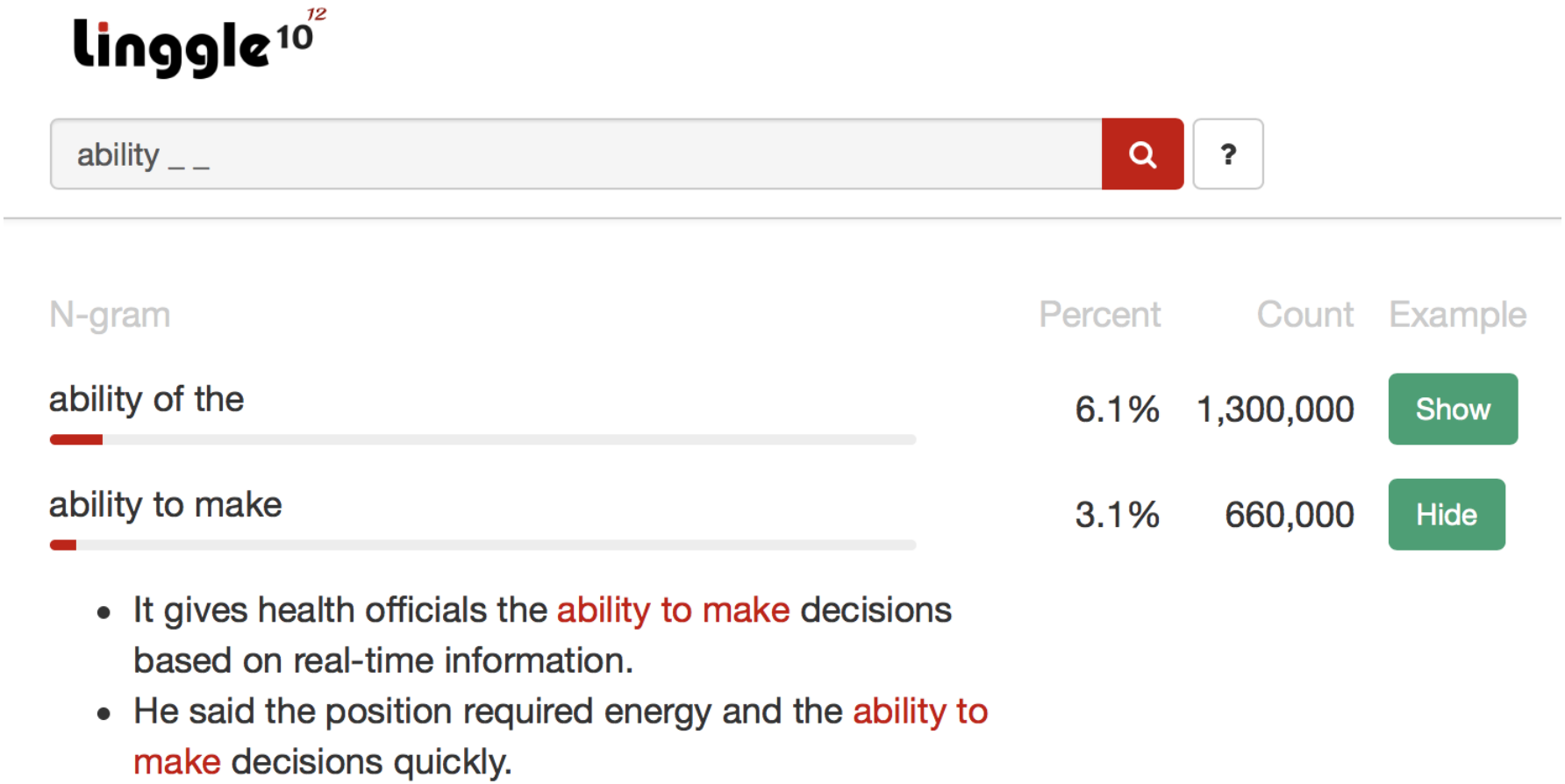


Use Linggle to Find Word Usage

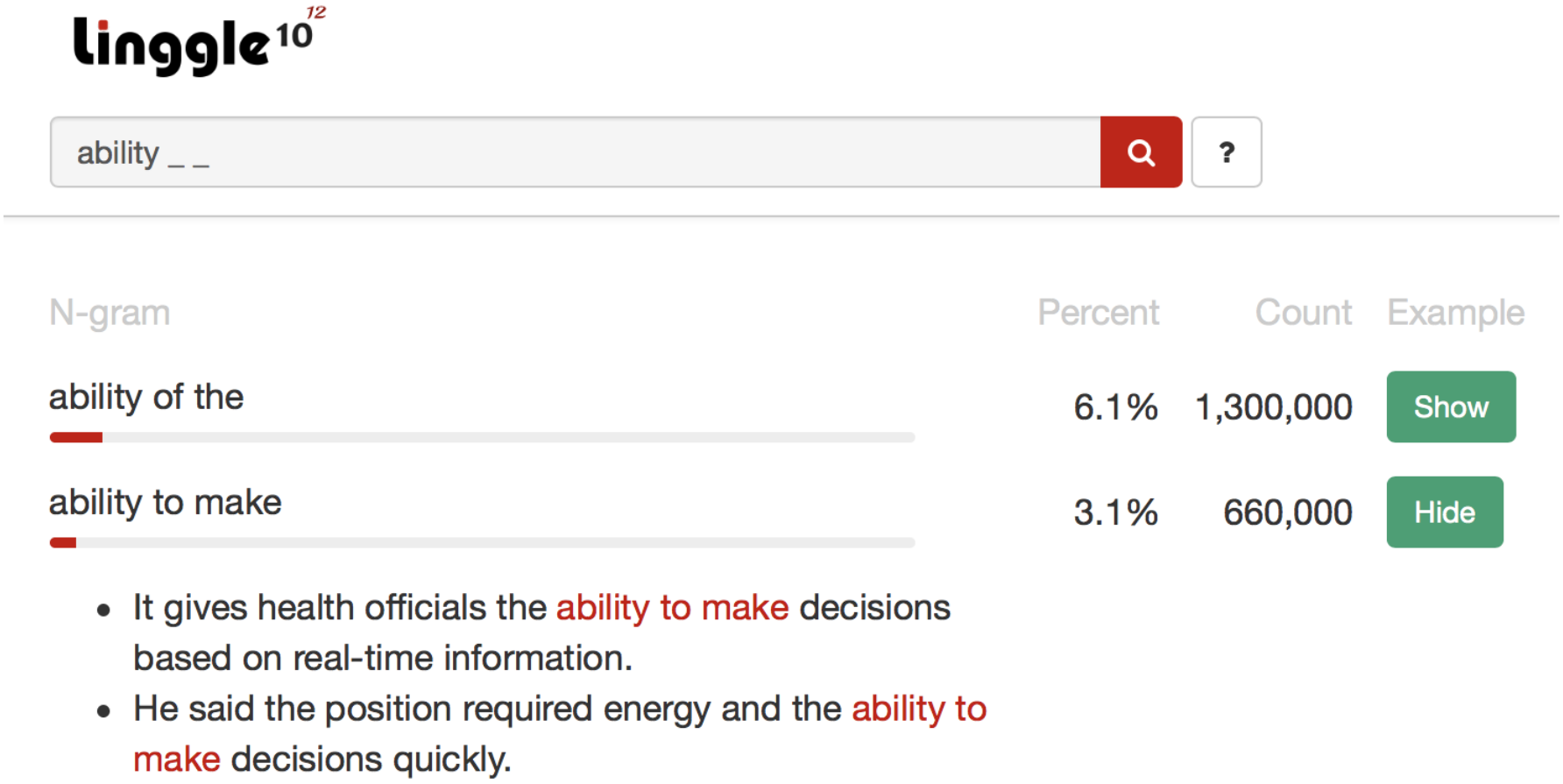


N-gram	Percent	Count	Example
ability of the	6.1%	1,300,000	Show
ability to make	3.1%	660,000	Show
ability to work	3.0%	640,000	Show
ability to use	2.8%	610,000	Show
ability to provide	2.3%	490,000	Show

Use Linggle to Find Word Usage with Examples

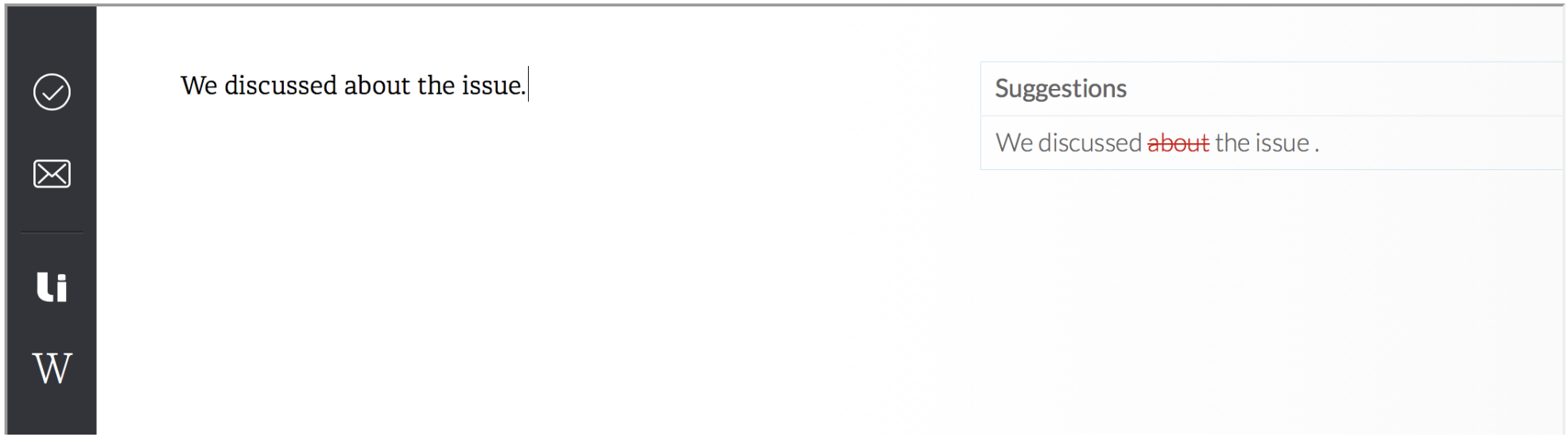


Linggle for COCA/ACADEMIC Is in the Work



開發文法檢查器—Cool English 試用版

文法偵錯系統



Source:
`https://www.coolenglish.edu.tw/moodle/mod/url/view.php?id=9735`

Questions?