



Multilingual Plagiarism Detection

Lalit Agarwal

Menna Mostafa

1 Problem Definition

According to Wikipedia, Plagiarism is:

- **Inserting** a text—copied word-for-word, or closely paraphrased with very few changes—from a source that is **not acknowledged** anywhere in the article.
- **Summarizing** a source in your own words, **without citing** the source in any way. [1]



2 Problem Definition

- In monolingual documents, plagiarism can be detected by means of document similarity, synonyms, ...
- What about Multilingual Plagiarism?

The look of Yosemite, from the toolbars to window construction, has been adjusted. Windows and the dock are now translucent.

Yosemite的外观，从工具栏到窗口设计，都进行了调整。窗口和停靠栏都是半透明的。



3 Challenges

- Machine translation (MT)?
 - There is more than one way to translate a word.
 - Different languages have different syntax.
 - Current MT tools are inefficient. [2]

The look of Yosemite, from the toolbars to window construction, has been adjusted. Windows and the **dock** are now translucent.

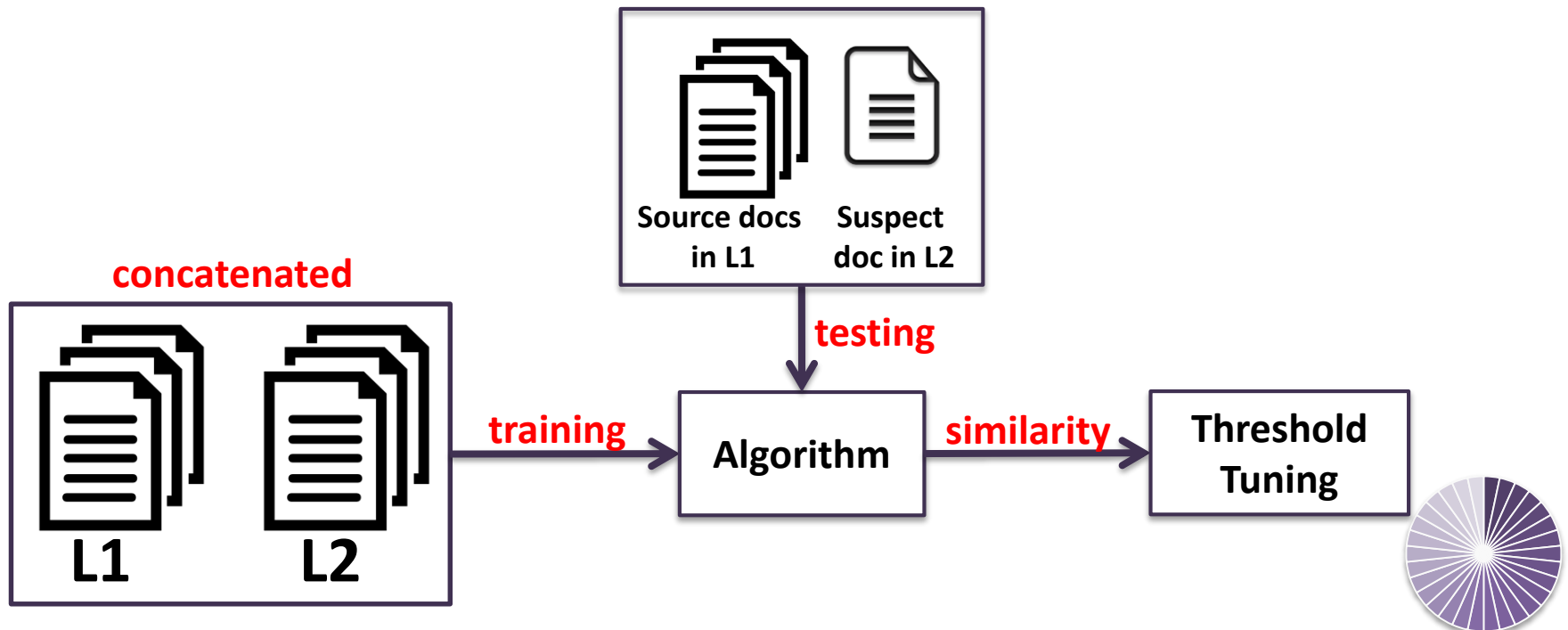
优山美地的外观，从工具栏窗口建设，进行了调整。
Windows和码头现在是半透明的。

Yosemite's appearance, from the toolbar window construction, have been adjusted. Windows and **Marina** is now translucent.



4 Approach

- Collection based Algorithms
 - Training with thousands of **parallel** or **comparable** documents about different topics.



5 Approach

- **Latent Semantic Indexing Algorithm(LSI):**
 - Based on: words that are used in the **same contexts** tend to have **similar meanings**. [3]
 - Given a document d_{L1} , LSI retrieves documents in L2 that are **conceptually similar** in meaning to d_{L1} even if the results **don't share a specific word** or words with d_{L1} . [3]



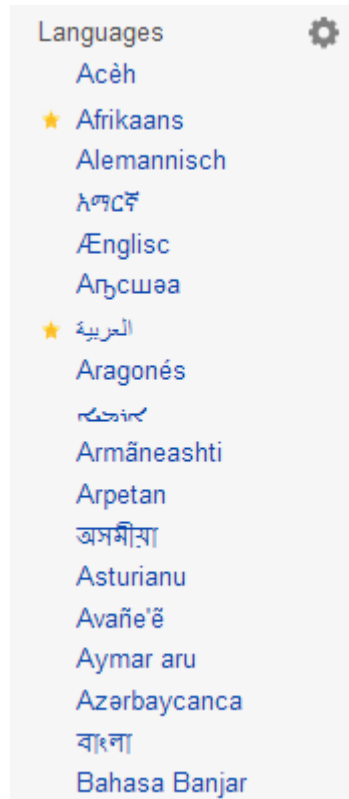
6 Approach

- The disadvantages of using collection based approach are:
 - Collection (training) data coverage.
 - Computationally and memory intensive.
- We will attempt to overcome some of these disadvantages by looping humans in.



7 Implementation

- **Language Pair:**
 - English <-> Chinese
- **Training Datasets:**
 - United Nations manually translated sessions **[Parallel]**. (politics, economic, human rights topics)
 - Wikipedia dataset using multilingual links and Wikipedia official dumps **[Comparable]**. (politics, science, sports, biographies, art, ...)
 - Mix of both



8 Implementation

- **LSI implementation:**
 - Used gensim (Python library) implementation of LSI.
- **Test Documents:**
 - 28 English-Chinese pairs collected manually from various sources such are:
 - Wikipedia Chinese featured documents.[4]
 - Edinburg University plagiarism dataset (translated by MT tools and verified by Jerry!).[5]
 - Xinhuanet bilingual zone. [6]
 - Yeeyan manually translated articles.[7]



9 Implementation

- **Documents Preprocessing:**
 - English documents:
 - Lower case
 - Clear non-alpha characters
 - Remove stop words.[8]
 - Stemming.[9]
 - Chinese documents:
 - Cut words (segmentation).[10]
 - Remove stop words.[8]
- Removed documents shorter than **30** words.
- Removed words that appeared in less than **5** documents.
- Removed words that appeared in more than **half** the documents.
- Afterwards, retained only the top **150,000** words from both languages.



10 Implementation

- LSI infers some latent topics.

- Example:

– Topic 1: $-0.184 * \text{"委员会"} + -0.170 * \text{"会议"} + 0.162 * \text{"安全"} + 0.148 * \text{"council"} + 0.139 * \text{"我们"} + 0.136 * \text{"secur"} + -0.135 * \text{"committe"} + -0.124 * \text{"a"} + -0.123 * \text{"大会"} + 0.115 * \text{"理事会"}$

committee

conference

security

general
assembly

council



11 Implementation

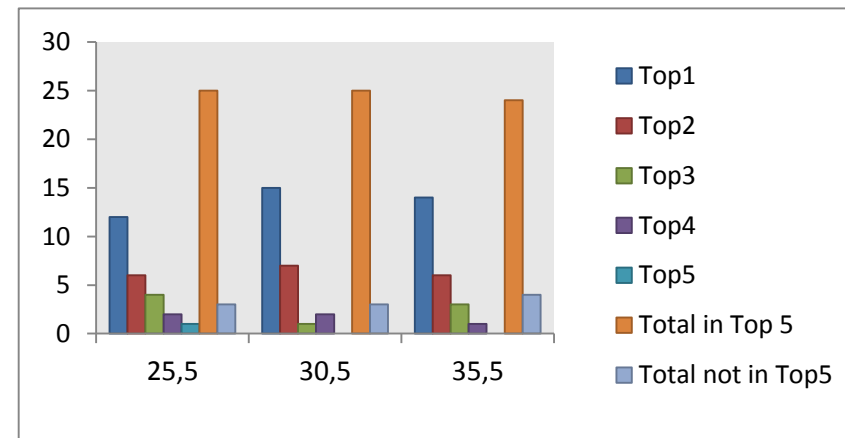
- Topics in the parallel test dataset:
 - Politics
 - The 1996 United States campaign finance controversy was an alleged effort by the People's Republic of China to influence domestic American politics
 - Economics
 - Bank of China (Hong Kong) Limited is the second-largest commercial banking group
 - Technology
 - Many people blame Microsoft's predicament on Steve Ballmer, the big, bald, manic, fist-pumping sales
 - Sports
 - Top seed Novak Djokovic is one match win away from becoming year-end world No. 1



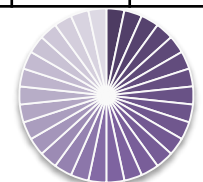
12 Results

Tested the system using the manually collected dataset:

- **28 Chinese** documents versus a collection of around **10000 English** documents (including corresponding pairs of the 28 docs).
- Some documents were not retrieved at all.
 - Top seed **Novak Djokovic** is one match win away from becoming year-end world No. 1.....
 - UN Secretary-General Ban Ki-moon said Friday that there is hope that the **ebola** outbreak could be contained by mid-2015

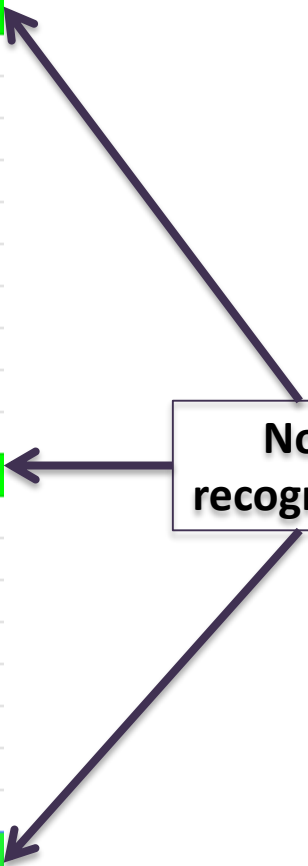


Dataset Size Wikipedia (in k),UN (in k)	Top1	Top2	Top3	Top4	Top5	Total in Top 5	Total not in Top5
25,5	12	6	4	2	1	25	3
30,5	15	7	1	2	0	25	3
35,5	14	6	3	1	0	24	4



doc ID	top1	top2	top3	top4	top5	Not in top 5
9995	10002:0.244318	9995:0.241621	9521:0.227007	2971:0.225436	5565:0.218973	
9996	9996:0.42459	7380:0.318104	8017:0.286381	3834:0.2861	700:0.285542	
9997	7428:0.272027	9323:0.199209	4203:0.160149	8016:0.159793	7453:0.159406	
9998	9998:0.15379	3060:0.149815	5057:0.134144	4561:0.133826	2947:0.123357	
9999	9999:0.326629	6010:0.178842	7426:0.176884	10001:0.166236	3241:0.1656	
10000	10000:0.286095	195:0.160096	10009:0.156029	6702:0.148463	6615:0.142424	
10001	10001:0.232291	5298:0.156159	5252:0.154189	6127:0.133772	7312:0.129184	
10002	10002:0.244318	9995:0.241621	9521:0.227007	2971:0.225436	5565:0.218973	
10003	7831:0.150514	10003:0.150514	7847:0.145938	2832:0.135866	6727:0.132127	
10004	9251:0.26151	2635:0.206752	10004:0.203041	7967:0.203041	2905:0.202943	
10005	5871:0.202647	10005:0.202647	621:0.157082	8366:0.155484	6882:0.145546	
10006	10006:0.431898	7071:0.431898	9287:0.431766	1995:0.365436	7031:0.354422	
10007	3297:0.215907	10007:0.215907	8762:0.188679	7755:0.180947	5017:0.172974	
10008	9426:0.281064	1109:0.264178	9038:0.222037	8900:0.185158	9986:0.173137	
10009	10009:0.40259	761:0.325885	3515:0.251154	10015:0.244404	577:0.206303	
10010	10010:0.506566	8424:0.345125	3237:0.274772	5283:0.253996	8325:0.252478	
10011	1905:0.250049	10011:0.237531	5854:0.235739	5343:0.225055	9930:0.2231	
10012	10012:0.283682	8508:0.241707	1574:0.219177	7230:0.20921	6952:0.198783	
10013	10013:0.28181	8278:0.275479	1853:0.264668	4580:0.248308	7434:0.244814	
10014	6074:0.459326	10014:0.433736	6245:0.358228	8361:0.353778	1425:0.346055	
10015	5821:0.370207	1947:0.338291	6841:0.331168	10015:0.300781	3182:0.291361	
10016	10016:0.221447	1420:0.202973	4323:0.202973	851:0.190042	3421:0.188941	
10017	9038:0.275893	1109:0.248936	9426:0.248851	2794:0.238646	2867:0.229935	
7831	7831:0.150514	10003:0.150514	7847:0.145938	2832:0.135866	6727:0.132127	
7967	9251:0.26151	2635:0.206752	10004:0.203041	7967:0.203041	2905:0.202943	
5871	5871:0.192901	10005:0.192901	2060:0.128025	1698:0.124953	275:0.120014	
7071	10006:0.431898	7071:0.431898	9287:0.431766	1995:0.365436	7031:0.354422	
3297	3297:0.215907	10007:0.215907	8762:0.188679	7755:0.180947	5017:0.172974	
Tops:	15	7	1	2	0	3

Not recognized



13 Results

- Some topics were **not included (or not well covered)** in the training documents, and in result they are unrecognized by the algorithm.
- Therefore, we ran **two experiments** with humans:
 1. Evaluate how users perceived plagiarism by displaying 4 articles, 2 plagiarized and 2 similar, and asking users for their opinions.
 2. Evaluate the algorithm similarity index and missing topics. (no one offered to help 😞)



15 Results

- Feedback received from users on Exp1 :
 - Reduce documents length.
 - Split documents to sentences.
 - Results were somewhat inconsistent.
- Each user took around 2-3 mins per document pair.

Users	Document:1 ISIS Same Author	Document:2 Ebola Same Author	Document:3 Online Privacy Diff Authors	Document:4 iPhone Security Diff Authors	Document:5 Exactly Translated Document
User1	Partially Plagiarized	Plagiarized	Partially Plagiarized	Not Plagiarized	Plagiarized
User2	Partially Plagiarized	Plagiarized	Not Plagiarized	Not Plagiarized	Not Plagiarized
User3	Partially Plagiarized	Plagiarized	Partially Plagiarized	Not Plagiarized	Partially Plagiarized
User4	Plagiarized	Plagiarized	Not Plagiarized	Partially Plagiarized	Not Plagiarized



14 Results (UI from Exp2)

CHINESE

帕洛阿尔托网络公司(Palo Alto Network)报告称,该公司发现了一种名为WireLurker的针对苹果移动设备及台式电脑的恶意软件,并称“这是我们见过的规模最大的恶意软件”。虽然这款恶意软件——旨在造成损害或盗取信息的软件——针对的是中国的用户,而且能够避免,但此次行动展示了攻击者侵袭装有苹果iOS系统的移动设备的新方式。该公司称,用户如果通过USB连接线将移动设备与Mac电脑连接,用户的iOS设备也会受到感染。“任何iOS设备只要通过USB连接到受感染的OS X电脑,并安装下载的第三方应用程序,或自动在设备上产生恶意应用程序,都会被WireLurker监控,不管设备是否已经越狱,”该公司安全研究人员说。“因此我们称之为‘wire lurker’(连接线中的潜藏者)。”研究人员称,一旦WireLurker被安装到Mac电脑上,这款恶意软件就会等待用户通过USB连接iOS设备,然后立即感染该设备。一旦被感染,WireLurker的制造者就能窃取受害人的通讯簿、读取iMessage中的短信并定期从攻击者的指挥控制服务器发出更新请求。尽管尚不清楚制造者的最终目的,但研究人员称,有人正在积极更新该恶意软件。

ENGLISH

Top1 Document Top2 Document Top3 Document

US-based Palo Alto Networks said WireLurker appeared to have originated in China and was mostly infecting devices there. The malware first targets Mac computers via a third-party store before copying itself to iOS devices. Researchers warn it steals information and can install other damaging apps. WireLurker has the ability to transfer from Apple's Mac computer to mobile devices through a USB cable. The malware initially gets onto an iOS device via a USB link to an infected Mac computer. The security firm said the malware was capable of stealing "a variety of information" from mobile devices it infects and regularly requested updates from the attackers' control server. According to Palo Alto Networks, WireLurker was first noticed in June when a developer at the Chinese firm Tencent realised there were suspicious files and processes happening on his Mac and iPhone. Further inquiries revealed a total of 467 Mac programs

Documents Similarity

0.23

FEEDBACK

How do you rate the performance of the algorithm?

- Very Good
- Good
- Bad

Do you think the Top1 document is plagiarised?

- Yes
- No
- I am not sure

If any, what are the topics which were present in the Chinese Doc but not in the English doc?

Topics

NEXT

Remove all Highlights

Start Over



16 Future Work

- Use the collected data from users to:
 - Tune the threshold for deciding if a document pair is plagiarized or not.
 - Add missing topics to the training data.
- Sample documents from Wikipedia that gives good and equal coverage over the general topics.



17 References

- [1] <http://en.wikipedia.org/wiki/Wikipedia:Plagiarism>
- [2] <https://www.plagiarismtoday.com/2011/02/24/the-problem-with-detecting-translated-plagiarism/>
- [3] http://en.wikipedia.org/wiki/Latent_semantic_indexing
- [4] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_China/Featured_and_good_content
- [5] <http://www.inf.ed.ac.uk/teaching/courses/tts/assessed/assessment3.html>
- [6] <http://www.xinhuanet.com/english/bilingual/news.htm>
- [7] <http://www.yeeyan.org/>
- [8] <https://code.google.com/p/stop-words/>
- [9] http://www.nltk.org/_modules/nltk/stem/porter.html
- [10] <http://code.google.com/p/smallseg/>



Question 1



Question 2



Thank You!