Using bitext mining to identify translated material:

practical assessment and new applications

Zhilu Tu, Minghao Wang,
Mark Shuttleworth, and Zhiwen Hua
Hong Kong Baptist University

TIIS @HKBU

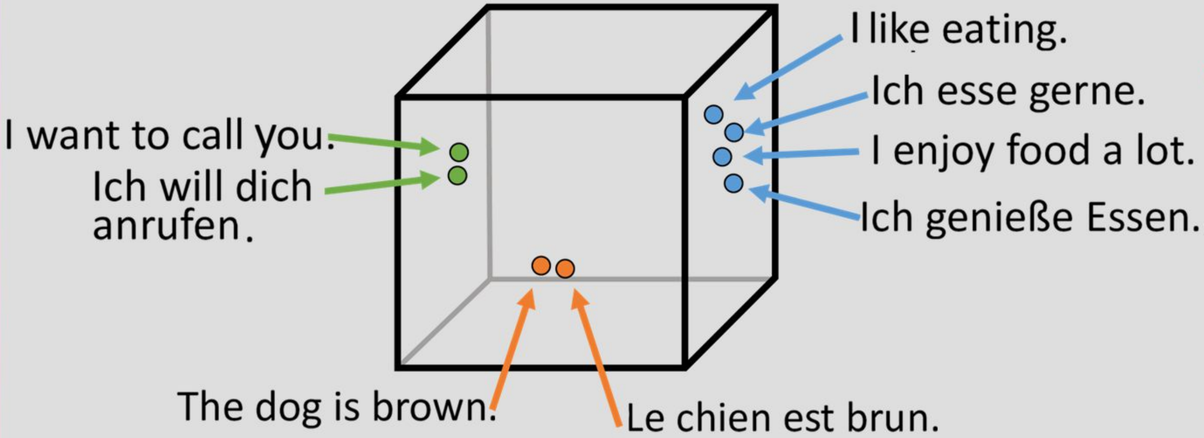
WikiAligner

Introduction

Bilingual text mining

SENTENCE LENGTH

LEXICON RECOURSES

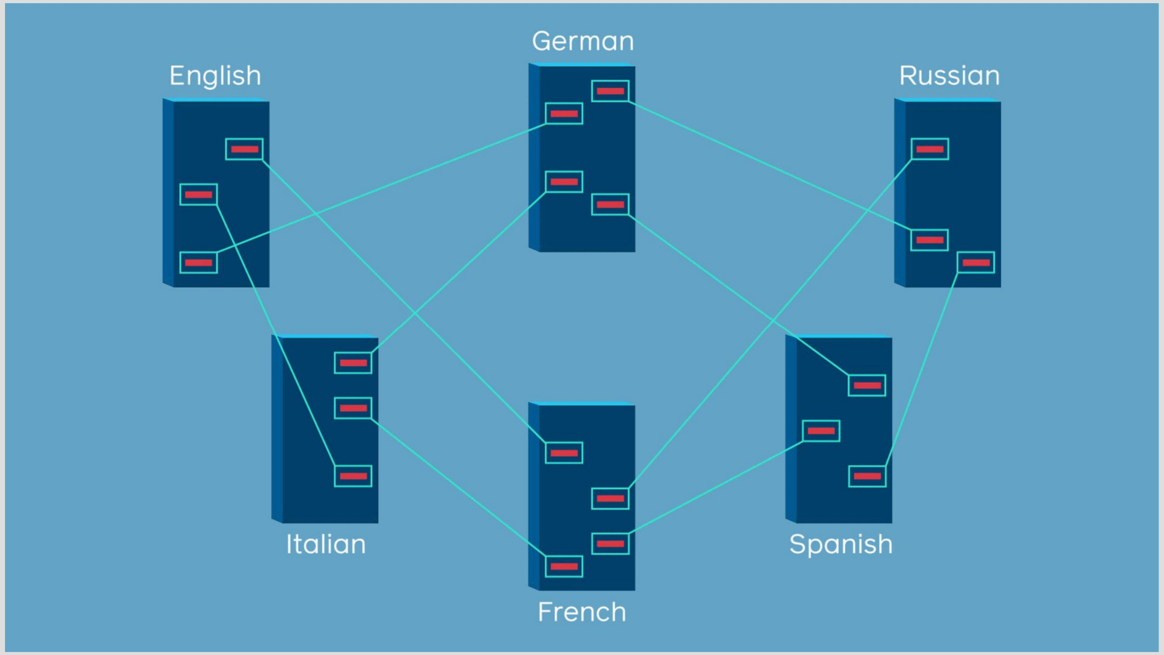


(Schwenk, 2019; Thompson & Zhang, 2019)

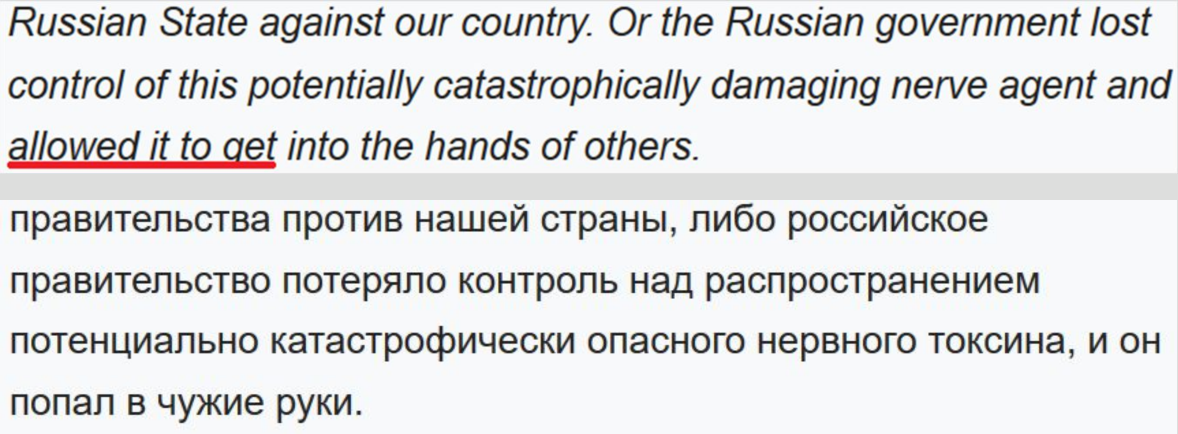
- Traditional
  - Sentence length-based algorithm
  - Bilingual cross-reference resources
  - Monotonic - can’t locate scattered sentences
- Conventional
  - Sentence-embedding
  - Pre-trained model – LASER; LaBSE
  - Non-monotonic

Objectives

Towards practicality



CCMatrix (Joulin & Schwenk, 2020)



Russian State against our country. Or the Russian government lost control of this potentially catastrophically damaging nerve agent and allowed it to get into the hands of others.

правительства против нашей страны, либо российское правительство потеряло контроль над распространением потенциально катастрофически опасного нервного токсина, и он попал в чужие руки.

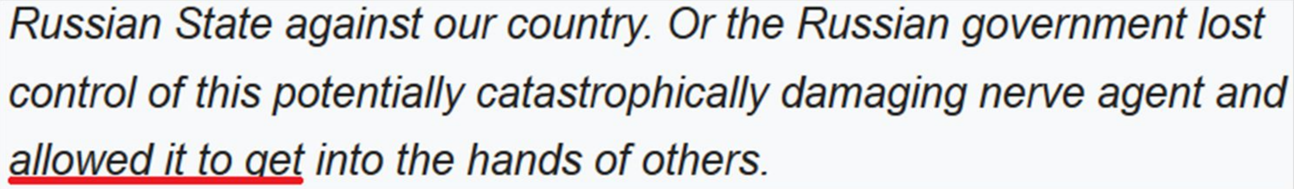
- A variety of big data applications have now been created using bitext mining technology.

- The paper investigates the use of bitext mining within a small-scale dataset setting as an example of how they might be exploited in other practical scenarios

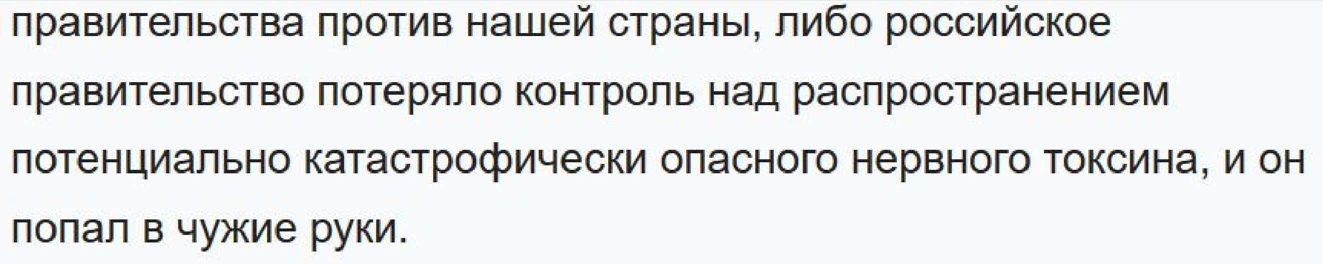
Objectives

Example

- Translation largely remains Wikipedia’s “dark matter”; this has led to discrepancies and changes in points of view (Shuttleworth, 2017).



Russian State against our country. Or the Russian government lost control of this potentially catastrophically damaging nerve agent and allowed it to get into the hands of others.

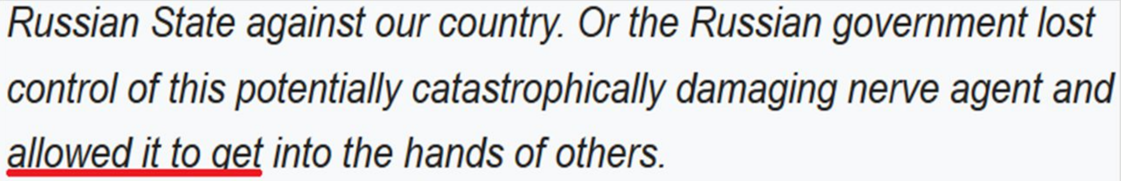


правительства против нашей страны, либо российское правительство потеряло контроль над распространением потенциально катастрофически опасного нервного токсина, и он попал в чужие руки.

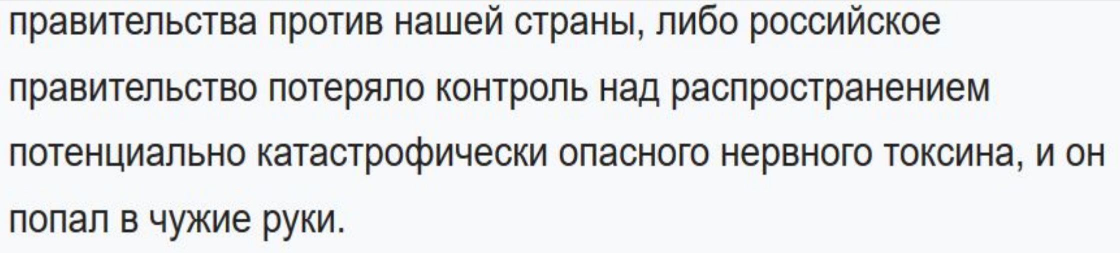
[... or the Russian government lost control of the distribution of a potentially catastrophically dangerous nerve toxin and it got into the hands of others]

Objectives

Hypothesis



Russian State against our country. Or the Russian government lost control of this potentially catastrophically damaging nerve agent and allowed it to get into the hands of others.

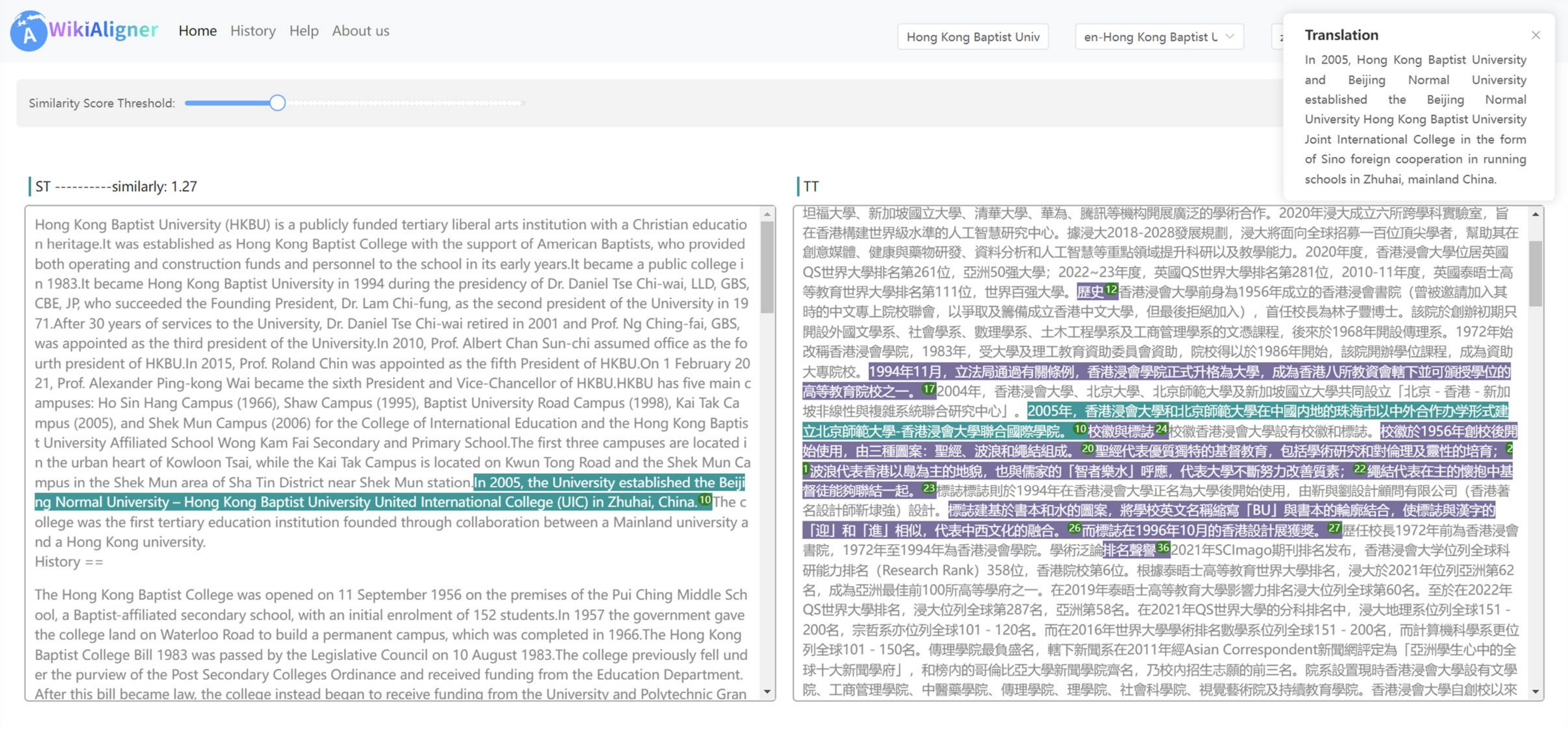


правительства против нашей страны, либо российское правительство потеряло контроль над распространением потенциально катастрофически опасного нервного токсина, и он попал в чужие руки.

[... or the Russian government lost control of the distribution of a potentially catastrophically dangerous nerve toxin and it got into the hands of others]

- Aligning materials manually affected the efficiency of the research and the possibility of enlarging the study scale

- Bitext mining can help reduce the researcher's workload while opening up the possibility of analyzing more data



Hong Kong Baptist Univ

en-Hong Kong Baptist U

Translation

In 2005, Hong Kong Baptist University and Beijing Normal University established the Beijing Normal University Hong Kong Baptist University Joint International College in the form of Sino foreign cooperation in running schools in Zhuhai, mainland China.

Similarity Score Threshold:

ST ----------similarly: 1.27

Hong Kong Baptist University (HKBU) is a publicly funded tertiary liberal arts institution with a Christian education heritage.It was established as Hong Kong Baptist College with the support of American Baptists, who provided both operating and construction funds and personnel to the school in its early years.It became a public college in 1983.It became Hong Kong Baptist University in 1994 during the presidency of Dr. Daniel Tse Chi-wai, LLD, GBS, CBE, JP, who succeeded the Founding President, Dr. Lam Chi-fung, as the second president of the University in 1971.After 30 years of services to the University, Dr. Daniel Tse Chi-wai retired in 2001 and Prof. Ng Ching-fai, GBS, was appointed as the third president of the University.In 2010, Prof. Albert Chan Sun-chi assumed office as the fourth president of HKBU.In 2015, Prof. Roland Chin was appointed as the fifth President of HKBU.On 1 February 2021, Prof. Alexander Ping-kong Wai became the sixth President and Vice-Chancellor of HKBU.HKBU has five main campuses: Ho Sin Hang Campus (1966), Shaw Campus (1995), Baptist University Road Campus (1998), Kai Tak Campus (2005), and Shek Mun Campus (2006) for the College of International Education and the Hong Kong Baptist University Affiliated School Wong Kam Fai Secondary and Primary School.The first three campuses are located in the urban heart of Kowloon Tsai, while the Kai Tak Campus is located on Kwun Tong Road and the Shek Mun Campus in the Shek Mun area of Sha Tin District near Shek Mun station.In 2005, the University established the Beijing Normal University – Hong Kong Baptist University United International College (UIC) in Zhuhai, China.[10]The college was the first tertiary education institution founded through collaboration between a Mainland university and a Hong Kong university.
History ==

The Hong Kong Baptist College was opened on 11 September 1956 on the premises of the Pui Ching Middle School, a Baptist-affiliated secondary school, with an initial enrolment of 152 students.In 1957 the government gave the college land on Waterloo Road to build a permanent campus, which was completed in 1966.The Hong Kong Baptist College Bill 1983 was passed by the Legislative Council on 10 August 1983.The college previously fell under the purview of the Post Secondary Colleges Ordinance and received funding from the Education Department. After this bill became law, the college instead began to receive funding from the University and Polytechnic Gran

TT

坦福大學、新加坡國立大學、清華大學、華為、騰訊等機构開展廣泛的學術合作。2020年浸大成立六所跨學科實驗室，旨在香港構建世界級水準的人工智慧研究中心。據浸大2018-2028發展規劃，浸大將面向全球招募一百位頂尖學者，幫助其在創意媒體、健康與藥物研發、資料分析和人工智慧等重點領域提升科研以及教學能力。2020年度，香港浸會大學位居英國QS世界大學排名第261位，亞洲50强大學；2022~23年度，英國QS世界大學排名第281位，2010-11年度，英國泰晤士高等教育世界大學排名第111位，世界百强大學。歷史[12]香港浸會大學前身為1956年成立的香港浸會書院（曾被邀請加入其時的中文專上院校聯會，以爭取及籌備成立香港中文大學，但最後拒絕加入），首任校長為林子豐博士。該院於創辦初期只開設外國文學系、社會學系、數理學系、土木工程學系及工商管理學系的文憑課程，後來於1968年開設傳理系。1972年始改稱香港浸會學院，1983年，受大學及理工教育資助委員會資助，院校得以於1986年開始，該院開辦學位課程，成為資助大專院校。1994年11月，立法局通過有關條例，香港浸會學院正式升格為大學，成為香港八所教資會轄下並可頒授學位的高等教育院校之一。[17]2004年，香港浸會大學、北京大學、北京師範大學及新加坡國立大學共同設立「北京 - 香港 - 新加坡非線性與複雜系統聯合研究中心」。2005年，香港浸會大學和北京師範大學在中國内地的珠海市以中外合作办学形式建立北京師範大學-香港浸會大學聯合國際學院。[10]校徽與標誌[24]校徽香港浸會大學設有校徽和標誌。校徽於1956年創校後開始使用，由三種圖案：聖經、波浪和繩結組成。[20]聖經代表優質獨特的基督教育，包括學術研究和對倫理及靈性的培育；[21]波浪代表香港以島為主的地貌，也與儒家的「智者樂水」呼應，代表大學不斷努力改善質素；[22]繩結代表在主的懷抱中基督徒能夠聯結一起。[23]標誌標誌則於1994年在香港浸會大學正名為大學後開始使用，由靳與劉設計顧問有限公司（香港著名設計師靳埭強）設計。標誌建基於書本和水的圖案，將學校英文名稱縮寫「BU」與書本的輪廓結合，使標誌與漢字的「迎」和「進」相似，代表中西文化的融合。[26]而標誌在1996年10月的香港設計展獲獎。[27]歷任校長1972年前為香港浸會書院，1972年至1994年為香港浸會學院。學術泛論排名聲譽[36]2021年SCImago期刊排名发布，香港浸會大学位列全球科研能力排名（Research Rank）358位，香港院校第6位。根據泰晤士高等教育世界大學排名，浸大於2021年位列亞洲第62名，成為亞洲最佳前100所高等學府之一。在2019年泰晤士高等教育大學影響力排名浸大位列全球第60名。至於在2022年QS世界大學排名，浸大位列全球第287名，亞洲第58名。在2021年QS世界大學的分科排名中，浸大地理系位列全球151 - 200名，宗哲系亦位列全球101 - 120名。而在2016年世界大學學術排名數學系位列全球151 - 200名，而計算機科學系更位列全球101 - 150名。傳理學院最負盛名，轄下新聞系在2011年經Asian Correspondent新聞網評定為「亞洲學生心中的全球十大新聞學府」，和榜内的哥倫比亞大學新聞學院齊名，乃校内招生志願的前三名。院系設置現時香港浸會大學設有文學院、工商管理學院、中醫藥學院、傳理學院、理學院、社會科學院、視覺藝術院及持續教育學院。香港浸會大學自創校以來

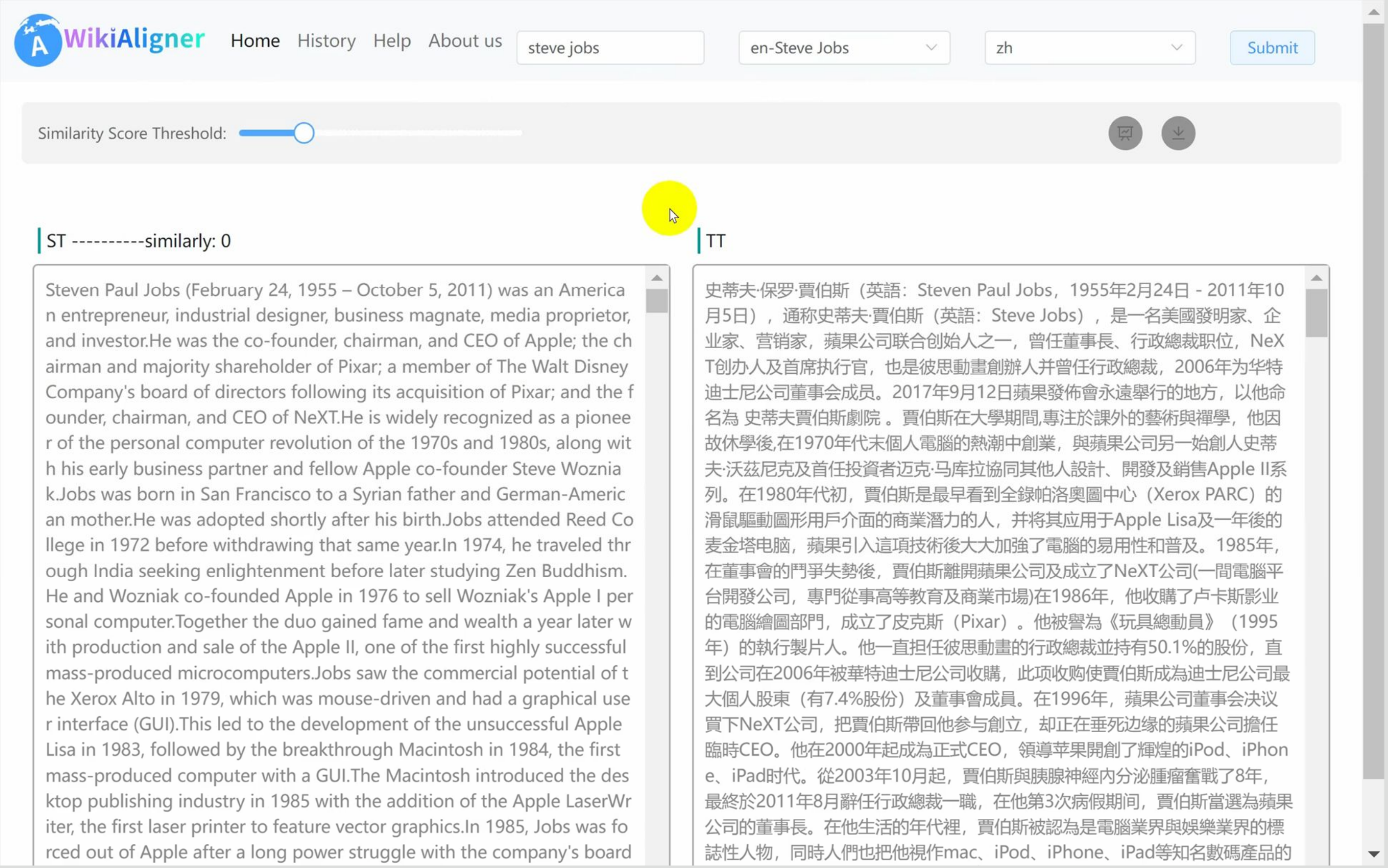
Methodology

Our tool

- WikiAligner: A Browser/Server Wikipedia bitext mining solution integrating back-end pipeline and front-end re-presentation of the data.

Methodology

Demo - Minghao Wang



steve jobs

en-Steve Jobs

zh

Similarity Score Threshold:

ST ----------similarly: 0

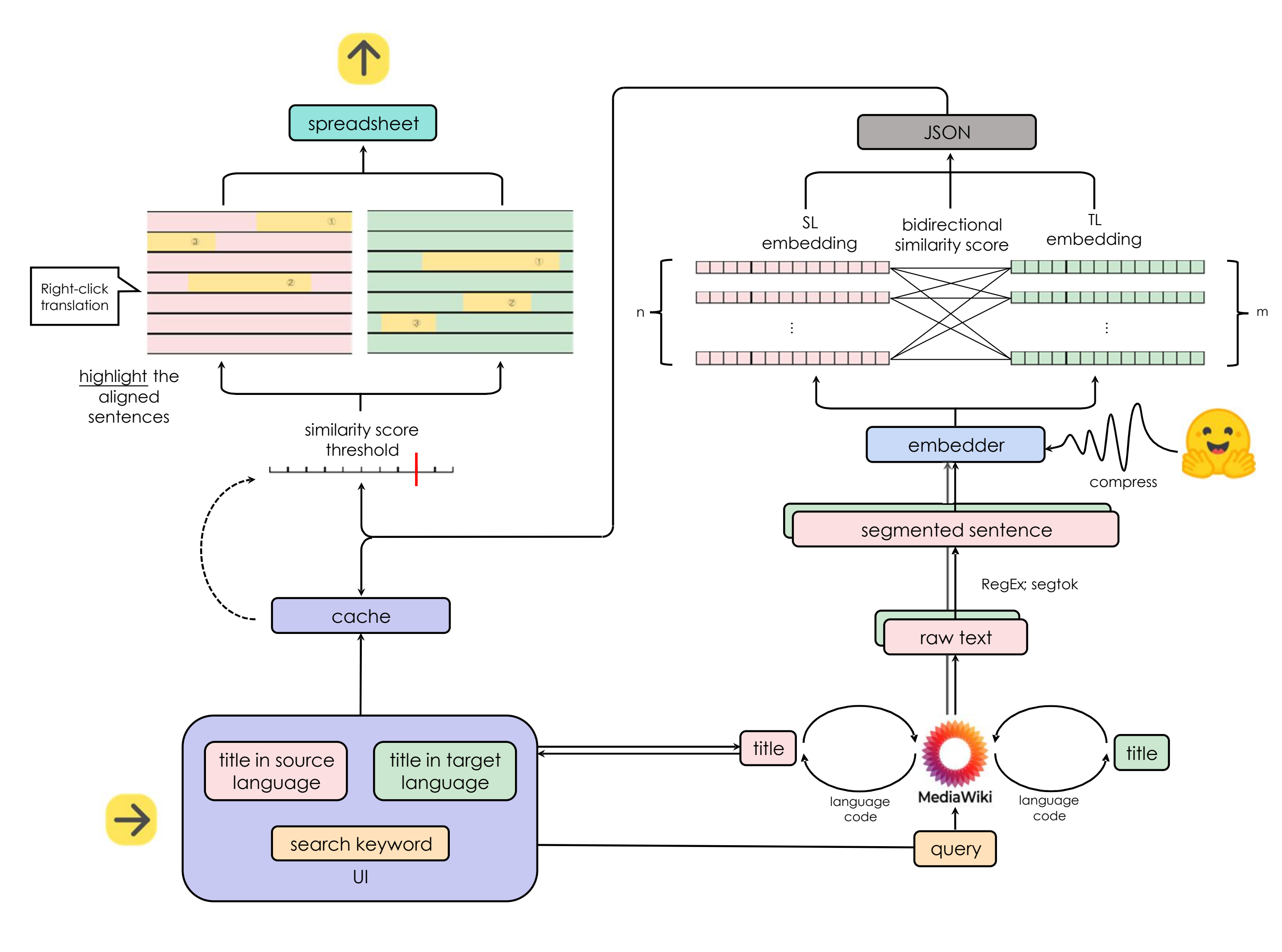
Steven Paul Jobs (February 24, 1955 – October 5, 2011) was an American entrepreneur, industrial designer, business magnate, media proprietor, and investor.He was the co-founder, chairman, and CEO of Apple; the chairman and majority shareholder of Pixar; a member of The Walt Disney Company's board of directors following its acquisition of Pixar; and the founder, chairman, and CEO of NeXT.He is widely recognized as a pioneer of the personal computer revolution of the 1970s and 1980s, along with his early business partner and fellow Apple co-founder Steve Wozniak.Jobs was born in San Francisco to a Syrian father and German-American mother.He was adopted shortly after his birth.Jobs attended Reed College in 1972 before withdrawing that same year.In 1974, he traveled through India seeking enlightenment before later studying Zen Buddhism. He and Wozniak co-founded Apple in 1976 to sell Wozniak's Apple I personal computer.Together the duo gained fame and wealth a year later with production and sale of the Apple II, one of the first highly successful mass-produced microcomputers.Jobs saw the commercial potential of the Xerox Alto in 1979, which was mouse-driven and had a graphical user interface (GUI).This led to the development of the unsuccessful Apple Lisa in 1983, followed by the breakthrough Macintosh in 1984, the first mass-produced computer with a GUI.The Macintosh introduced the desktop publishing industry in 1985 with the addition of the Apple LaserWriter, the first laser printer to feature vector graphics.In 1985, Jobs was forced out of Apple after a long power struggle with the company's board

TT

史蒂夫·保罗·賈伯斯（英語：Steven Paul Jobs，1955年2月24日 - 2011年10月5日），通称史蒂夫·賈伯斯（英語：Steve Jobs），是一名美國發明家、企业家、营销家，蘋果公司联合创始人之一，曾任董事長、行政總裁职位，NeXT创办人及首席执行官，也是彼思動畫創辦人并曾任行政總裁，2006年为华特迪士尼公司董事会成员。2017年9月12日蘋果發佈會永遠舉行的地方，以他命名為 史蒂夫賈伯斯劇院 。賈伯斯在大學期間,專注於課外的藝術與禪學，他因故休學後,在1970年代末個人電腦的熱潮中創業，與蘋果公司另一始創人史蒂夫·沃兹尼克及首任投資者迈克·马库拉協同其他人設計、開發及銷售Apple II系列。在1980年代初，賈伯斯是最早看到全錄帕洛奧圖中心（Xerox PARC）的滑鼠驅動圖形用戶介面的商業潛力的人，并将其应用于Apple Lisa及一年後的麦金塔电脑，蘋果引入這項技術後大大加強了電腦的易用性和普及。1985年，在董事會的鬥爭失勢後，賈伯斯離開蘋果公司及成立了NeXT公司(一間電腦平台開發公司，專門從事高等教育及商業市場)在1986年，他收購了卢卡斯影业的電腦繪圖部門，成立了皮克斯（Pixar）。他被譽為《玩具總動員》（1995年）的執行製片人。他一直担任彼思動畫的行政總裁並持有50.1%的股份，直到公司在2006年被華特迪士尼公司收購，此项收购使賈伯斯成為迪士尼公司最大個人股東（有7.4%股份）及董事會成員。在1996年，蘋果公司董事会决议買下NeXT公司，把賈伯斯帶回他参与創立，却正在垂死边缘的蘋果公司擔任臨時CEO。他在2000年起成為正式CEO，領導苹果開創了輝煌的iPod、iPhone、iPad时代。從2003年10月起，賈伯斯與胰腺神經內分泌腫瘤奮戰了8年，最終於2011年8月辭任行政總裁一職，在他第3次病假期间，賈伯斯當選為蘋果公司的董事長。在他生活的年代裡，賈伯斯被認為是電腦業界與娛樂業界的標誌性人物，同時人們也把他視作mac、iPod、iPhone、iPad等知名數碼產品的

Methodology

WikiAligner workflow



1. Front-end UI input
   - Keywords for the articles
   - SL title & TL title
2. Back-end data pipeline
   - Segmented sentences
   - LaBSE(Feng et al., 2022) embedding
   - Similarity search
3. Front-end display
   - Highlight
   - Tags & Threshold
   - Right-click translation

Discussion

Assessments

WikiAligner

Performance

- All the alignments that have been extensively analyzed by Shuttleworth (2018) are also located by using our tool.

Accessibility

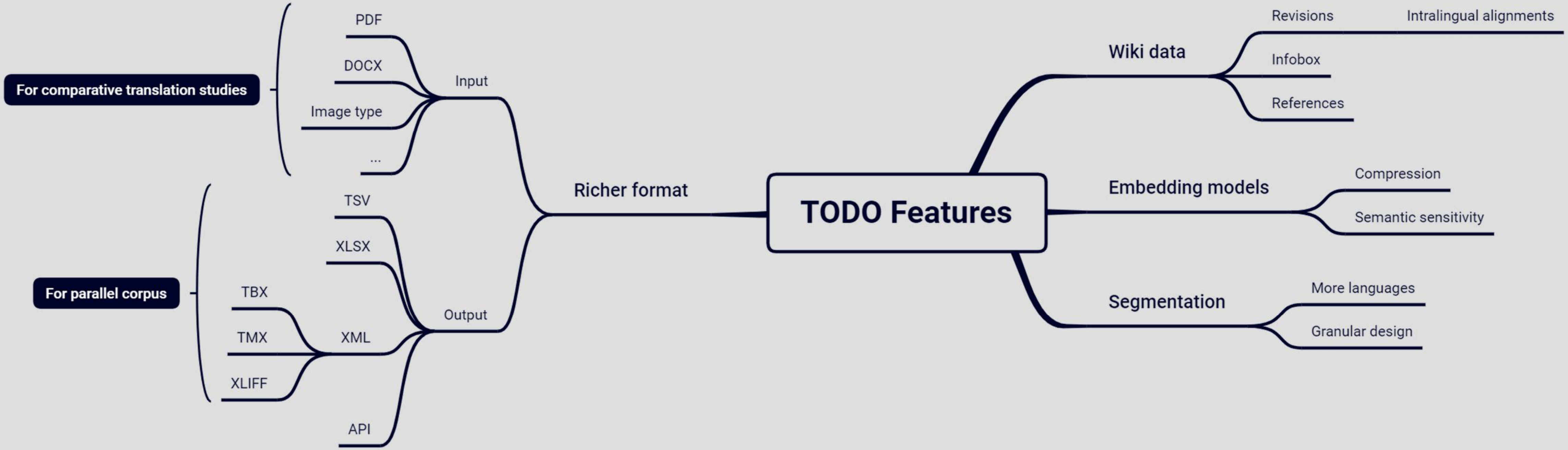
- As with other web-based services, potential users do not need to attend the coding of the bitext tool before using

Manifestations

- Researchers will be able to scrutinize every translation pair and their distribution by using the highlight feature

Discussion

New applications



Conclusion

- Standing on the shoulders of giants from NLP, we present a method to utilize bitext mining in the context of a small-scale dataset.
- WikiAligner can be an entry point for applying bitext mining to more real-world scenarios with new features.

References

- Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT Sentence Embedding. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 878–891. https://doi.org/10.18653/v1/2022.acl-long.62
- Joulin, A., & Schwenk, H. (2020, February). *CCMatrix: A billion-scale bitext dataset for training translation models*. https://ai.facebook.com/blog/ccmatrix-a-billion-scale-bitext-data-set-for-training-translation-models/
- Schwenk, H. (2019, January 22). *LASER natural language processing toolkit - Engineering at Meta*. https://engineering.fb.com/2019/01/22/ai-research/laser-multilingual-sentence-embeddings/
- Shuttleworth, M. (2017). Locating foci of translation on Wikipedia. *Translation Spaces*, *6*(2), 310–332. https://doi.org/10.1075/TS.6.2.07SHU
- Shuttleworth, M. (2018). Translation and the Production of Knowledge in “Wikipedia”: Chronicling the Assassination of Boris Nemtsov. *Alif: Journal of Comparative Poetics*, *38*, 231–263. https://www.jstor.org/stable/26496376
- Thompson, B., & Zhang, S. (2019). *thompsonb/vecalign: Improved Sentence Alignment in Linear Time and Space*. https://github.com/thompsonb/vecalign

Q&A

- wikialigner@outlook.com