AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE   
WYDZIAŁ ELEKTROTECHNIKI, AUTOMATYKI, INFORMATYKI I ELEKTRONIKI

**System wykrywania podobieństw kodów źródłowych w projektach studenckich**

**PRACA MAGISTERSKA**

**Jarosław Szczęśniak**

**Promotor: dr inż. Darin Nikolow**

Kraków 2011

Oświadczamy, świadomi odpowiedzialności karnej za poświadczenie nieprawdy, że niniejszą pracę dyplomową wykonaliśmy osobiście i samodzielnie (w zakresie wyszczególnionym we wstępie) i że nie korzystaliśmy ze źródeł innych niż wymienione w pracy.

Spis treści

[1.Wstęp 4](#__RefHeading__2139_1984353016)

[1.1Kontekst pracy 4](#__RefHeading__2141_1984353016)

[1.2Nowe podejście do procesu tłumaczenia 4](#__RefHeading__2143_1984353016)

[1.3Struktura pracy 4](#__RefHeading__2145_1984353016)

[2.Podstawy systemów tłumaczeniowych 5](#__RefHeading__2147_1984353016)

[2.1Segmentacja tekstu 5](#__RefHeading__2149_1984353016)

[2.2Tłumaczenie maszynowe 5](#__RefHeading__2151_1984353016)

[2.2.1Silniki tłumaczenia maszynowego dostępne online 5](#__RefHeading__2153_1984353016)

[2.2.2Badanie efektywności silników MT 5](#__RefHeading__2155_1984353016)

[2.2.3Wielosilnikowe tłumaczenie maszynowe 5](#__RefHeading__2157_1984353016)

[2.3Tłumaczenie wspierane komputerowo 5](#__RefHeading__2159_1984353016)

[2.4Tłumaczenie przez jednojęzycznych tłumaczy 5](#__RefHeading__2161_1984353016)

[2.5Podsumowanie rozdziału 5](#__RefHeading__2163_1984353016)

[3.Tłumaczenie iteracyjne przez jednojęzycznych edytorów 6](#__RefHeading__2165_1984353016)

[4.System ITM – analiza wymagań 6](#__RefHeading__2167_1984353016)

[4.1Wymagania funkcjonalne wobec platformy ITM 6](#__RefHeading__2169_1984353016)

[4.2Rozszerzalność systemu ITM 6](#__RefHeading__2171_1984353016)

[4.3Wymagania funkcjonalne dla aplikacji do komunikacji IM 6](#__RefHeading__2173_1984353016)

[4.4Wymagania niefunkcjonalne dla komunikacji IM 6](#__RefHeading__2175_1984353016)

[4.5Wymagania funkcjonalne dla aplikacji do tłumaczenia dokumentów 6](#__RefHeading__2177_1984353016)

[4.6Wymagania niefunkcjonalne dla aplikacji do tłumaczenia dokumentów 6](#__RefHeading__2179_1984353016)

[5.Projekt i architektura systemu ITM 7](#__RefHeading__2181_1984353016)

[6.Implementacja systemu ITM 8](#__RefHeading__2183_1984353016)

[7.Badania eksperymentalne 9](#__RefHeading__2185_1984353016)

[8.Wnioski 9](#__RefHeading__2187_1984353016)

[Bibliografia 11](#__RefHeading__2189_1984353016)

[Spis Ilustracji 16](#__RefHeading__2191_1984353016)

1. Wstęp

Niniejsza praca prezentuje założenia tłumaczenia iteracyjnego przez osoby jednojęzyczne (ang. *Iterative Translation by Monoliguists – ITM*) oraz opisuje implementację i badania platformy usług ITM. ITM jest nowym podejściem do tłumaczenia tekstu, które w przeciwieństwie do tłumaczenia tradycyjnego, angażuje w ten proces osoby mogące znać jedynie swoje języki ojczyste. Model ITM pozwala na otrzymanie tłumaczenia o wyższej jakości niż w przypadku tłumaczenia maszynowego oraz w krótszym czasie niż w modelu tradycyjnym. Koncepcja ma szansę przynieść znaczącą redukcję bariery językowej podczas wielojęzycznego przepływu informacji oraz stać się istotnym elementem w procesie tworzenia niezawodnego silnika tłumaczenia maszynowego.

W ramach prac badawczych nad tą tematyką stworzono pilotażową platformę usług tłumaczenia iteracyjnego wraz z dedykowanymi aplikacjami do tłumaczenia tekstów i komunikacji natychmiastowej, działającymi w sieci Internet. Opracowano także ITM API umożliwiające wykorzystanie usług platformy przez inne aplikacje. Następnie przeprowadzone zostały badania skuteczności procesu ITM oraz przetestowano sam system pod względem używalności, wydajności i rozszerzalności.

* 1. Kontekst pracy

W dobie rozwoju Internetu jego użytkownicy uzyskali możliwość swobodnego dostępu do informacji oraz ich wymiany. Sieć WWW umożliwia ludziom przeglądanie najnowszych wiadomości, artykułów prasowych, publikacji naukowych oraz innych tekstów nie tylko w ich języku ojczystym, ale także w różnych językach obcych. Coraz większe znaczenie ma także bezpośrednia komunikacja pomiędzy użytkownikami z całego świata za pośrednictwem poczty elektronicznej i komunikatorów internetowych. W związku z tak rozumianą globalizacją w Internecie zachodzi potrzeba wielojęzycznej komunikacji w różnych kontekstach. Coraz większa ilość stron oraz portali internetowych jest tłumaczona na kilka języków, aby maksymalnie poszerzyć grono potencjalnych odbiorców. Zdarza się, że dbają o to również autorzy blogów, którym zależy na jak największej liczbie czytelników. Poza tym wciąż rośnie liczba osób, dla których istotna jest bezpośrednia komunikacja poprzez Internet z obcokrajowcami, zarówno na płaszczyźnie prywatnej jak i zawodowej.

* 1. Nowe podejście do procesu tłumaczenia
  2. Struktura pracy

1. Podstawy systemów tłumaczeniowych
   1. Segmentacja tekstu
   2. Tłumaczenie maszynowe
      1. Silniki tłumaczenia maszynowego dostępne *online*
      2. Badanie efektywności silników MT
      3. Wielosilnikowe tłumaczenie maszynowe
   3. Tłumaczenie wspierane komputerowo
   4. Tłumaczenie przez jednojęzycznych tłumaczy
   5. Podsumowanie rozdziału
2. Tłumaczenie iteracyjne przez jednojęzycznych edytorów

Rys. Model tłumaczenia tradycyjnego

Rys. Kooperacja edytora źródłowego i docelowego

1. System ITM – analiza wymagań
   1. Wymagania funkcjonalne wobec platformy ITM
   2. Rozszerzalność systemu ITM
   3. Wymagania funkcjonalne dla aplikacji do komunikacji IM
   4. Wymagania niefunkcjonalne dla komunikacji IM
   5. Wymagania funkcjonalne dla aplikacji do tłumaczenia dokumentów

Rys. Diagram przypadków użycia dla aplikacji do tłumaczenia tekstów

* 1. Wymagania niefunkcjonalne dla aplikacji do tłumaczenia dokumentów

1. Projekt i architektura systemu ITM

1. Implementacja systemu ITM

1. Badania eksperymentalne

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Własność/nr | I | II | III | IV | V | VI |
| Język źródłowy | angielski | angielski | angielski | polski | polski | francuski |
| Język docelowy | polski | polski | polski | angielski | angielski | angielski |
| Typ tekstu | powieść przygodowa | tekst techniczny | artykuł prasowy | tekst techniczny | artykuł prasowy | artykuł prasowy |
| Ilość segmentów | 33 | 45 | 28 | 20 | 28 | 10 |
| Ilość słów | 633 | 804 | 610 | 321 | 533 | 192 |

Tab. Charakterystyka tłumaczonych tekstów

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | I | II | III | IV | V | VI |
| BLEU (5-gramy) dla MT | 0.152 | 0.207 | 0.101 | 0.192 | 0.172 | 0.243 |
| BLEU (5-gramy) dla ITM | 0.301 | 0.356 | 0.221 | 0.341 | 0.316 | 0.381 |
| Poprawa | 98% | 71% | 118% | 77% | 83% | 56% |

Tab. Ocena tłumaczeń przez algorytm BLEU

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | I | II | III | IV | V | VI |
| Ocena dla MT | 2.2 / 2.0 | 3.1 / 2.1 | 2.9 / 2.0 | 3.2 / 2.9 | 3.0 / 1.9 | 3.3 / 3.1 |
| Ocena dla ITM | 3.0 / 3.1 | 4.2 / 4.2 | 3.9 / 3.5 | 4.3 / 4.0 | 3.8 / 3.2 | 4.1 / 4.1 |
| Poprawa | 36% / 55% | 28% / 52% | 34% / 75% | 43% / 38% | 26% / 68% | 24% / 32% |

Tab. Ocena adekwatności i poprawności językowej tłumaczenia dokonana przez człowieka

Rys. Czas odpowiedzi platformy w zależności od ilości zapytań

1. Wnioski

Bibliografia

1. **The Gallup Organization.** User language preferences online - Analytical report. 2011.

2. **Kay, Martin.** The Proper Place of Men and Machines in Language. 1980.

3. **Potępa, Anna, et al.** Iterative Translation by Monolinguists - tests of the new approach. *Lecture Notes in Computer Science.* 2011.

4. G**azeta.pl.** StartupFest. [Online] http://startupfest.pl.

5. **Nguyen, ThuyLinh, Vogel, Stephan and Smith, Noah A.** Nonparametric Word Segmentation for Machine Translation. 2010.

6. **Palmer, David D.** Tokenisation and sentence segmentation.

7. **International Organization for Standardization.** *ISO TC37/SC4.*

8. **Hutchins, John.** Machine translation: A concise history. 2007.

9. **Narjes Sharif Razavian, Stephan Vogel.** The Web as a Platform to Build Machine Translation. [Online] 2009.

10. **Callison-Burch, Chris, et al.** Findings of the 2010 JointWorkshop on Statistical Machine Translation and Metrics for Machine Translation. *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR.*

11. **iTranslate.** [Online] http://itranslate4.eu/.

12. **Yahoo.** Yahoo Babelfish. [Online] http://babelfish.yahoo.com/.

13. **Microsoft.** Microsoft Translator. [Online] http://www.microsofttranslator.com/.

14. **K. Papineni., S. Roukos, T. Ward, W. Zhu.** Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL.* 2002.

15. **Callison-Burch, Chris, et al.***Findings of the 2009 Workshop on Statistical Machine Translation.*

16. **Callison-Burch, Chris, et al.** (Meta-) Evaluation of Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation.* 2007.

17. **Callison-Burch, Chris, et al**. Further Meta-Evaluation of Machine Translation. *Proceedings of the Third Workshop on Statistical Machine Translation.* 2008.

18. **Savoy, Jacques and Dolamic, Ljiljana.** *How Effectiveis Google's Translation System in Search.*

19. **Nomoto, Tadashi.** Multi-Engine Machine Translation with Voted Language Model.

20. **Koehn, Philipp, et al.** Moses: Open Source Toolkit for Statistical Machine Translation. 2007.

21. **Chen, Yu, et al.** Combining Multi-Engine Translations with Moses. 2009.

22. **Barrachina, Sergio, et al.** Statistical Approaches to Computer-Assisted Translation. 2008.

23. **Daniel Ortiz-Martinez, Ismael Garcia-Varea, Francisco Casacuberta.** Online Learning for Interactive Statistical Machine Translation. [Online]

24. **Wu, Jian-Cheng, et al.** Subsentential Translation Memory for Computer Assisted Writing and Translation. 2004.

25. **XTRF.** [Online] http://pl.xtrf.eu/.

26. **Organization for the Advancement of Structured Information Standards.** XLIFF specification. [Online] http://www.oasis-open.org/committees/xliff/documents/cs-xliff-core-1.1-20031031.htm.

27. **Joshua S. Albrech, and Rebecca Hwa, G. Elisabeta Marai.** Correcting Automatic Translations through Collaborations between MT and Monolingual Target-Language Users. 2009 .

28. **Koehn, Philipp.** Enabling Monolingual Translators: Post-Editing vs. Options. [Online]

29. **Resnik, Philip, et al.** Improving Translation via Targeted Paraphrasing. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.*

30. **Tbot for Windows Live Messenger**. [Online] http://www.microsofttranslator.com/user/bot/.

31. **Translation bots for GTalk**. [Online] http://www.google.com/support/talk/bin/answer.py?answer=89921.

32. **Ignite Realtime.** Spark. [Online] http://www.igniterealtime.org/projects/spark/index.jsp.

33. **Localization Industry Standards Association.** http://www.lisa.org. [Online]

34. **Casacuberta, Francisco, et al.** Human Interaction For High-Quality Machine Translation. *Communications of the ACM.* 2009.

35. **Henderson, Cal.** Building Scalable Web Sites. 2006.

36. **Apache.** HBase. [Online] http://wiki.apache.org/hadoop/Hbase.

37. **Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach.** Bigtable: A Distributed Storage System for Structured Data. 2006.

38. **Pytel, M.** Rozproszone składowanie danych dla potrzeb systemu tłumaczenia iteracyjnego. 2011.

39. **XMPP Standards Foundation.** XMPP: Instant Messaging and Presence. [Online] http://xmpp.org/rfcs/rfc6121.html.

40. **XMPP Standards Foundation**. XMPP: Core. [Online] http://xmpp.org/rfcs/rfc6120.html.

41. **Ignite Realtime.** Smack API. [Online] www.igniterealtime.org/projects/smack/.

42. **Google .** Narzędzia Google dla Tłumaczy. [Online] http://translate.google.com/toolkit/.

43. **Google.** Google Translate API for Java. [Online] http://code.google.com/p/google-api-translate-java/.

44. **Google .** University Research Program for Google Translate. [Online] http://research.google.com/university/translate/.

45. **Okapi Framework.** [Online] http://okapi.sourceforge.net/.

46. **Springsource.** Spring. [Online] http://www.springsource.org/.

47. **World Wide Web Consortium.** WSDL. [Online] http://www.w3.org/TR/wsdl.

48. **World Wide Web Consortium**. SOAP. [Online] http://www.w3.org/TR/soap/.

49. **Apache.** Apache CXF. [Online] http://cxf.apache.org/.

50. **Ignite Realtime.** Sparkplug Kit. [Online] http://www.igniterealtime.org/projects/spark/sparkplug-kit.jsp.

51. **Google.** Google translate REST Api. [Online] http://code.google.com/intl/pl/apis/language/translate/v1/getting\_started.html.

52. **Apache.** Apache HttpClient. [Online] http://hc.apache.org/httpcomponents-client-ga/.

53. **Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach.** Bigtable: A Distributed Storage System for Structured Data. *OSDI 2006.*

54. **World Wide Web Consortium.** XPath. [Online] http://www.w3schools.com/xpath/.

55. **Apache.** Apache POI. [Online] http://poi.apache.org/.

56. **Milkowski, Marcin and Lipski, Jarosław.** *Using SRX standard for sentence segmentation in LanguageTool.* 2009.

57. **Google.** Google Docs. [Online] https://docs.google.com/.

58. **Matt, Raible.** Appfuse 2. [Online] http://appfuse.org/display/APF/Home.

59. **Direct Web Remoting**. [Online] http://directwebremoting.org.

60. **Google.** Google Wave. [Online] http://wave.google.com/.

61. **Google**. Google Gadget. [Online] http://code.google.com/apis/wave/extensions/gadgets/guide.html.

62. **Google**. Google Wave Robot. [Online] http://code.google.com/apis/wave/extensions/robots/.

63. **Apache.** JMeter. [Online] http://jakarta.apache.org/jmeter/.

64. **The 3rd Asian Conference on Intelligent Information and Database Systems**. [Online] http://ke.yu.ac.kr/aciids2011/.

65. **Google.** The University Research Program for Google Translate. [Online] http://research.google.com/university/translate/docs.html.

66. **Lopez, Adam.** Statistical machine translation. 2008.

67. **Workshop on statistical machine translation**. [Online] http://www.statmt.org/wmt11/.

68. **Sergio Barrachina, Oliver Bender, Francisco Casacuberta.** Statistical Approaches to Computer-Assisted Translation. 2008.

69. **Callison-Burch, Chris, Koehn, Philipp and Osborne, Miles.** Improved Statistical Machine Translation Using Paraphrases in Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. 2006.

70. **Madnani, Nitin, et al.** Using Paraphrases for Parameter Tuning in Statistical Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation.* 2007.

71. **Albrech, Joshua S., Hwa, Rebecca and Marai, G. Elisabeta.** Correcting Automatic Translations through Collaborations between MT and Monolingual Target-Language Users. 2009.

72. **Gamma, Erich, et al.** *Design Patterns - Elements of reusable Object Oriented Software.* Warszawa : Wydawnictwo Nukowo Techniczne, 2005.

73. **Gangadharaia, Rashmi, Brown, Ralf and Carbonell, Jaime.** Monolingual Distributional Profiles forWord Substitution in Machine Translation. *Coling 2010: Poster Volume.* 2010.

74. **Shinyama, Yusuke, Sekine, Satoshi and Sudo, Kiyoshi.** Automatic Paraphrase Acquisition from News Articles. *Proceedings of HLT 2002, Second International Conference on Human Language Technology Research.* 2002.

75. **Yahoo.** Yahoo Babelfish. [Online] http://babelfish.yahoo.com/.

76. **Google.** Google Translate. [Online] http://translate.google.com.

77. **Localization Industry Standards Association.** Segmentation Rules Exchange. [Online] http://www.lisa.org/fileadmin/standards/srx20.html.

78. **Wordpress.** [Online] http://wordpress.com/.

79. **Translators without borders.** [Online] http://tsf.eurotexte.fr/?lang=en.

80. **World Wide Web Consortium.** SOAP. [Online] http://www.w3.org/TR/soap/.

81. **Java Content Repository.** [Online] http://www.day.com/specs/jcr/1.0/.

82. **iTranslate4.eu.** [Online] http://iTranslate4.eu.

Spis Ilustracji