وزارت علوم تحقیقات و فناوری

گروه کامپیوتر

پایان نامه برای دریافت درجه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی

طراحی سامانه‌ای برای شناسایی جملات مشابه برای استفاده در خلاصه‌سازی چند سنده

استاد راهنما

محمد رضا فیضی درخشی

استاد مشاور

کریم صمد زمینی

پژوهش‌گر

باقر نوجوان آقدرق

سال دفاع

۱۳۹۲

چکیده

با افزایش روزافزون اطلاعات و حجم بالای مطالب موجود در فضای مجازی دیگر تکنیک‌های بازیابی اطلاعات جوابگوی نیاز کاربران نمی‌باشد. لزوم استفاده از روش‌هایی که بتواند خلاصه‌ای از اطلاعات مفید را در اختیار کاربران قرار دهد افزایش می‌یابد. خلاصه‌ساز چند سنده از جمله سیستم‌هایی می‌باشد که با ورود چند سند با موضوع یکسان به عنوان ورودی خلاصه‌ای از مهمترین اطلاعات مورد نظر را در اختیار کاربر قرار می‌دهد. افزونگی اطلاعات یکی از چالش‌های مهم در خلاصه‌سازی چند سنده می‌باشد. منظور از افزونگی اطلاعات تکرار مفاهیم یکسان در موضوع مشخص می‌باشد. با این توضیح که با افزونگی اطلاعات فرصت قرارگیری جملات مفید در خلاصه با توجه به حجم محدود آن از دست می‌رود. لذا لزوم دستیابی به سیستم شباهت‌یابی که بتواند از تکرار جملات مشابه در خلاصه جلوگیری کند افزایش می‌یابد.

روش شباهت‌یابی پیشنهادی در این پایان‌نامه که برای زبان فارسی طراحی گردیده بر پایه معنا و نحو می‌باشد که بعد از پیش‌پردازش و ریشه‌یابی و استخراج کلمات صورت می‌گیرد. در روش شباهت‌یابی بر پایه معنا از یک بردار یکتا که از کلمات دو جمله تشکیل شده بهره می‌بریم. سپس با استفاده از فارس‌نت که شامل کلمات، مترادفات و روابط موجود بین مترادفات کلمات می‌باشد برای دستیابی به شباهت دو کلمه و کامل کردن درایه‌های بردار یکتا برای هر جمله استفاده می‌کنیم. همچنین در روش شباهت‌یابی بر پایه معنا از برچسب‌گذار ادات سخن برای ارزش‌گذاری به نقش‌های مختلف کلمات (اسم، فعل و صفت) بهره می‌بریم. در روش شباهت‌یابی بر پایه نحو از ترتیب کلمات بهره برده‌ایم که به جایگاه قرارگیری کلمات با توجه به شباهت معنایی بین آن‌ها در جمله توجه می‌نماید. در نهایت با ضریبی که به شباهت معنا و نحو با توجه به اهمیت آن‌ها اختصاص داده می‌شود شباهت دو جمله محاسبه می‌گردد.

برای ارزیابی سیستم شباهت‌یاب از یک خلاصه‌ساز چند سنده بهره برده‌ایم. در این سیستم خلاصه‌ساز، از روش خوشه‌بندی Average Link و گزینشگری استفاده نموده‌ایم که شبیه‌ترین جمله در هر خوشه را انتخاب می‌کند. با بررسی آزمایش‌های به‌دست آمده و مقایسه روش پیشنهادی با روش شباهت‌یابی که در سیستم MEAD استفاده شده بود با بهبود ۷ درصدی در کاهش افزونگی مواجه شدیم.

کلمات کلیدی: خلاصه‌سازی چند سنده – افزونگی اطلاعات – شباهت‌یابی جملات

فهرست مطالب

۱. اخوان،ت ، شمس فرد،م ، عرفانی جورابچی،م ، "خلاصه‌ساز تک سنده و چند سنده متون فارسی: PARSUMIST "، چهاردهمین کنفرانس انجمن کامپیوتر ایران، تهران، ۱۳۸۷.

۲. مشکی،م ، آنالویی،م ، " خلاصه سازی چند سنده متون فارسی با استفاده ازیک روش مبتنی بر خوشه‌بندی"، اولین کنفرانس ملی مهندسی نرم افزار ایران، رودهن، ۱۳۸۸.

3. Chen,f., Chen,w., "research of sentence similarity computation method based on the enhanced petri net", *computational intelligence and software engineering*, 2009 .

۴. تشکری،م ، میبدی،م ، "طراحی یک ریشه‌یاب خودکار برای واژگان فارسی"، هفتمین کنفرانس سالانه انجمن کامپیوتر ایران، اسفند ۱۳۸۰.

۵. قاسم‌ثانی،غ ، حسامی‌فرد،ر ،"طراحی یک الگوریتم ریشه‌یابی برای زبان فارسی"، *یازدهمین کنفرانس بین‌المللی کامپیوتر انجمن کامپیوتر ایران*، تهران، بهمن ۱۳۸۴.

۶. نوجوان آقدرق،ب ، رمضانی،م ، فیضی درخشی،م،ر ، " ریشه‌یابی خودکار واژگان زبان فارسی با استفاده از ترکیبی بهینه از قوانین ساخت‌واژه و پایگاه‌داده‌ها"، *هشتمین همایش بین المللی انجمن ترویج زبان و ادب فارسی ایران*، زنجان، ۱۳۹۲.

7. Liu,X,Y., Zhou,Y,M., Zheng,R,S., "measuring semantic similarity within sentences", *proceeding of the seventh international conference on machine learning and cybernetics*, 2008.

8. Li,Y., Mclean,D., Bandar,Z,A., Oshea,J,D., Crockett,K., "sentence similarity based on semantic nets and corpus statistics", *IEEE Transactions on knowledge and data engineering* , 2006.

9. Lee,M,C., "a novel sentence similarity measure for semantic-based expert systems", *expert systems with application*, 2011.

۱۰. ستوده،آ، پویان،ع ، ستوده،ح ، "استخراج شباهت معنایی به‌وسیله سیستم استنتاج فازی در

خلاصه‌سازی متن"، *کنفرانس ملی فناوری اطلاعات و جهاد اقتصادی*، کازرون، ۱۳۹۰.

11. Lee,M,C., Zhang,J,W., Lee,W,X., Ye,H,Y., "sentence similarity computation based on pos and semantic nets", *fifth international joint conference on INC,IMS and IDC*, 2009.

12. Abdalgader,K., Skaber,A., "short text similarity measurement using word sense disambiguation and synonym expansion", *advances in artificial intelligence*, 2010.

13. Li,Y., Bandar,Z,A., Mclean,D., "an approach for measuring semantic similarity between words using multiple information sources", *IEEE Transaction on knowledge and data engineering* , 2003.

14. Bhagwani,S., Karnick,H., "sranjans: Semantic Textual Similarity using Maximal Weighted Bipartite Graph Matching", *First Joint Conference on Lexical and Computational*, 2012.

15. Olive,J., Serrano,J,L., Castillo,M,D,D., Lglesias,A., "SYMSS: A syntax-based measure for short-text semantic similarity", *data & knowledge engineering* , 2011.

16. Xiaoying,L., Yiming,Z., "sentence similarity based on dynamic time warping", *international conference on semantic computing*, 2007.

17. Chang,J,w., lee,M,c., wang,T,i., su,C,y., hsieh,T,c., "using grammar patterns to evaluate semantic similarity for short texts", *computing technology and information management(ICCM), 8th international conference on*, 2012.

18. Li,Y., Bandar,Z., mclean,D., Oshea,J., "a method for measuring sentence similarity and its application to conversational agents", *FLAIRS Conference*, 2004.

19. Selvi,P., gopalan,N,p., "sentence similarity computation based on wordnet and corpus statistics", *international conference on computational intelligence and multimedia applications*, 2007.

20. Li,L., Zhou,Y., Yuan,B., Wang,J., Hu,X., "sentence similarity measurement based on shallow parsing", *sixth international conference on fuzzy systems and*

knowledge discovery, 2009.

21. Tian,Y., Li,H., Cai,Q., Zhao,S., "measuring the similarity of short texts by word similarity and tree kernels", *information computing and telecommunications*, 2010.

22. Li,L., Hu,X., Hu,B,Y., Wang,J., Zhou,Y,M., "measuring sentence similarity from different aspects", *proceeding of the eighth international conference on machine learning and cybernetics*, 2009.

23. Shahabi,A,S., Kangavari,M,R., "a fuzzy approach for Persian text segmentation based on semantic similarity of sentences", *intelligent information processing* , 2007.

24. Skabar,A., Abdalgader,K., "improving sentence similarity measurement by incorporating sentential word importance", *advances in artificial intelligence*, 2011.

25. Ho,C., Murad,M,A,A., Doraisamy,S,C., Kadir,R,A., "measuring sentence similarity from both the perspective of commonalities and differences", *22nd international conference on tools with artificial intelligence* , 2010.

26. Banerjee,S., Pedersen,T., "extended gloss overlaps as a measure of semantic relatedness", *international joint conference on artificial intelligence* , 2003.

27. Zhu,T., Li,K., "The Similarity Measure Based on LDA for Automatic Summarization", *International Workshop on Information and Electronics Engineering*, 2012.

28. Sun,Y., Park,S,C., "Generation of Non-redundant Summary Based on Sum of Similarity", *Information Technology: Coding and Computing, ITCC. International Conference on*, 2005.

29. Feng,J., Zhou,Y., Martin,T., "Sentence Similarity based on Relevance", *Proceedings of IPMU'08, Torremolinos (M_alaga),* 2008.

30. Jian-fang,S., Zong-tian,L., Wen,Z., "Sentence Similarity Measure Based on Events and ContentWords", *Fuzzy systems and Knowledge Discovery, FSKD '09. Sixth International Conference on* , 2009 .

31. Poormasoomi,A., Kahani,M., Yazdi,S,V., Kamyar,H., "Context-Based Persian Multi-Document Summarization (global view)", *International Conference on Asian Language Processing*, 2011.

32. Achananuparp,P., Hu,X,H., Shen,X., "the evaluation of sentence similarity measures", *data warehousing and knowledge discovery*, 2008 .

۳۳. مشکی،م ، "*خلاصه‌سازی گزینشی چندسندی زبان فارسی*"، پایان‌نامه کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، اردیبهشت ۱۳۸۸.

34. Handl,J., Knowles,J., Dorigo,M., " Ant-based clustering: a comparative study of its relative performance with respect to K-means, average link and 1d-som", *Proceedings of the Third International Conference on Hybrid Intelligent Systems*, *IOS Press*, 2003.

۳۵. مشکی،م ، آنالویی،م ، "خلاصه‌سازی چند سندی متون فارسی با استفاده از یک روش مبتنی بر خوشه‌بندی"، *اولین کنفرانس ملی مهندسی نرم افزار ایران* ، رودهن ۱۳۸۸.