Simplificación Léxica para la Mejora de la Accesibilidad Cognitiva

Lexical Simplification to Improve Cognitive Accessibility

Rodrigo Alarcon[1]
[1] Grupo HULAT, Departamento Informática
Universidad Carlos III de Madrid
ralarcon@inf.uc3m.es

Resumen: Los textos que contienen palabras inusuales pueden provocar barreras de accesibilidad en las personas con discapacidad intelectual debido a la dificultad encontrada en la lectura y en la comprensión del texto. Como espacio de solución, se presenta esta propuesta de Tesis Doctoral, que incluye un enfoque para la simplificación léxica del texto que proporciona herramientas de accesibilidad como un soporte sistemático al cumplimento de estándares de accesibilidad. El enfoque se ha concretizado en el sistema de simplificación léxica EASIER el cual aborda también la tarea de la desambiguación del sentido de las palabras. En esta propuesta, se han aplicado diferentes técnicas de aprendizaje automático y embeddings contextuales (BERT) utilizando recursos de lectura fácil y lenguaje simple, como diccionarios y corpora. Esta contribución está orientada a ser una ayuda para las personas con discapacidad cognitiva en el acceso a la información en un dominio genérico y en idioma español.

Palabras clave: Discapacidad cognitiva, Simplificación léxica, Accesibilidad, Identificación de Palabras Complejas, Sinónimos, Definiciones, Desambiguación del sentido de las palabras

Abstract: Texts which contain unusual words can create accessibility barriers for individuals with intellectual disabilities due to the high degree of difficulty experienced when reading said words and the overall understanding of the text. With this motivation, this Doctoral Thesis proposal is presented, which includes an approach to lexical simplification that provides accessibility tools that systematically support compliance with accessibility standards. The approach has been concretized in the EASIER lexical simplification system, which also addresses the task of word-sense disambiguation. In this approach, different machine learning techniques and contextual embeddings (BERT) have been applied using Easy Reading and Simple Language resources such as dictionaries and corpora. The purpose of this contribution is to promote access to information found in generic domains in Spanish for individuals with cognitive disabilities.

Keywords: Cognitive disability, Lexical simplification, Accessibility, Complex Word Identification, Synonyms, Definitions, Word-sense disambiguation

1 Justificación de la investigación propuesta

Tenemos acceso a una cantidad abrumadora de información, pero esta información no es accesible para todas las personas. Algunas personas se enfrentan con barreras de accesibilidad cuando leen textos que contienen oraciones largas, palabras inusuales, estructuras lingüísticas complejas, etc.

Aunque las personas con discapacidad cognitiva del lenguaje y del aprendizaje se ven directamente afectadas, las barreras de accesibilidad cognitiva afectan a otros grupos de usuarios como las personas sordas, sordo-ciegas, mayores, analfabetos e inmigrantes con un idioma nativo diferente. Según el informe PISA del 2013, en España la mayoría de la población

Proceedings of the Doctoral Symposium on Natural Language Processing from the PLN.net network (PLNnet-DS-2020)

adulta tiene dificultades para entender textos densos (OCDE, 2013). Además, el 1,7% de la población es analfabeta funcional y hay 277.472 personas con una discapacidad intelectual.

Con el fin de proporcionar acceso universal a la información y hacer que los textos sean más accesibles, existen iniciativas que trabajan en la mejora de la accesibilidad cognitiva al lenguaje. Podemos destacar las pautas de Lectura Fácil (UNE, 2018), las cuales proponen pautas para adaptar textos que permitan una lectura y una comprensión más sencilla. Además, existe la iniciativa de Lenguaje Sencillo que promueve un lenguaje sencillo en el contenido de la sociedad de la información (EU, 2011).

Las Pautas de Accesibilidad al Contenido en la Web (WCAG) (W3C, 2019), incluyen pautas específicas para la mejora de la comprensión de los textos. En esta línea, otra iniciativa destacable es el Grupo de trabajo de accesibilidad para discapacidades cognitivas y de aprendizaje (COGA TF) (W3C, 2020) del W3C.

En todos estos trabajos se repite como esencial la pauta de utilizar un lenguaje con un léxico sencillo. Las técnicas de la pauta 3.1.3 (palabras inusuales) de las WCAG 2.1 recomienda ofrecer definiciones de las palabras inusuales y distinguirlas para que los usuarios puedan reconocerlas. El seguimiento de estas pautas ha definido los objetivos de esta propuesta de investigación: crear mecanismos para la detección de palabras complejas, ofreciendo de estas, sinónimos sencillos y definiciones en el ámbito de la accesibilidad.

Con esta motivación, surge esta propuesta de investigación que tiene como objetivo principal brindar un soporte sistemático al seguimiento de pautas de accesibilidad cognitiva usando métodos de Procesamiento de Lenguaje Natural (PLN) como la simplificación léxica y la desambiguación del sentido de las palabras. Ante la desventaja de no haber recursos en el ámbito de la accesibilidad cognitiva en español, se tiene como objetivo intermedio crear recursos del lenguaje en el ámbito de la accesibilidad cognitiva.

2 Origen y trabajo relacionado

2.1 Simplificación léxica

La simplificación léxica identifica palabras complejas y tiene la tarea de encontrar la mejor sustitución candidata para esas palabras objetivo. Si bien existen diferentes formas de lograr la simplificación léxica, Shardlow (Shardlow, 2014) divide el proceso de simplificación en cuatro pasos: Identificación de Palabras Complejas (CWI), Generación de Sustitutos, Selección de Sustitutos y Ranking de Sustitutos. En esta propuesta de investigación se utiliza el enfoque de Shardlow.

En la tarea de CWI los enfoques de aprendizaje automático (ML) han demostrado superar otras estrategias. Trabajos en esta línea se puede encontrar en SemEval 2016 (Task: Complex Word Identification) (Paetzold & Specia, 2016) y en el Workshop BEA 2018 (Yimam et al., 2017). En trabajos más recientes como en (Cheng, 2019) se utiliza una Red Neuronal Convolucional (CNN) junto con Word Embeddings y características lingüísticas/ morfológicas obteniendo óptimos resultados.

En relación con el segundo paso, la generación de sustitutos, la mayoría de los trabajos pueden agruparse en dos estrategias: consulta de bases de datos lingüísticas y la generación automática (Paetzold & Specia, 2017). Si bien esta estrategia tiene la ventaja de presentar un enfoque muy preciso, también tiene la desventaja de no tener una cobertura amplia. Otro enfoque encontrado es usando una Base de Datos de Paráfrasis multilingüe (PPDB) (Pavlick & Callison-Burch, 2016).

En el tercer paso, la selección de un sustituto del conjunto de los sinónimos extraídos en el paso anterior, se selecciona el sinónimo más adecuado según factores como la sencillez y su contexto. En esta etapa, el sinónimo seleccionado debe preservar el significado original de la oración, así como una correcta estructura sintáctica. Destacar (Paetzold & Specia, 2015) en el que el sinónimo final utiliza similitud semántica y modelos de Word Embedding.

2.2 Desambiguación del sentido de las palabras (WSD)

El lenguaje humano es ambiguo, muchas palabras pueden ser interpretadas de múltiples maneras dependiendo del contexto. Dado que nuevas palabras siguen siendo agregadas a nuestro lenguaje, esta tarea se vuelve cada vez más compleja y debido a que influye el dominio en el que el conocimiento es creado, la producción de recursos para apoyar a WSD resulta muy costoso. Teniendo en cuenta esta

desventaja, se pueden encontrar trabajos desde enfoques basados en conocimiento (Lesk, 1986), supervisados (Moradi, Ansari, & Zabokrtský, 2019), no supervisados (Chen, Bowes, & Brown, 2009), o semi supervisados (Cao, Bai, & Shinnou, 2019), e incluso competiciones (Navigli, Jurgens, & Vannella, 2013) (Moro & Navigli, 2015), las cuales tratan de afrontar ese problema de distintas maneras.

Sin embargo, otros afrontaron este problema desde otro punto de vista. Tal es el caso de Google, que utiliza su modelo de representación del lenguaje BERT (Bidireccional Encoder Representations from Transformers) (Devlin et al., 2019). En (Du, Qi, & Sun, 2019) se refinó BERT para la tarea de WSD, utilizando WordPiece embeddings como parte de las entradas, y obteniendo buenos resultados al sobrepasar los resultados de los enfoques actuales en F1-score.

2.3 PLN y discapacidad

Se encuentran trabajos de simplificación de texto para proporcionar soluciones tecnológicas a personas con discapacidad cognitiva como (Inui et al., 2003), que llevó a cabo la simplificación de texto proporcionando parafraseo sintáctico y léxico de un texto para ayudar a comprender el significado del texto en personas sordas. En (Carroll et al., 1998) se aborda la simplificación léxica para evitar vocabulario inusual con el que las personas con afasia podrían tener problemas. Destacar los proyectos Simplext[1] y FIRST[2] que presentan simplificación de texto en español para personas con discapacidad intelectual (Saggion et al., 2015) y para personas con autismo (Barbu, et al., 2015), respectivamente.

Se han encontrado trabajos que incluyen la tarea de WSD en sus sistemas de simplificación como en (Medina et al., 2016). En beneficio de las personas con discapacidad con problemas de comunicación y del lenguaje, se han encontrado trabajos de sistemas que incluyen WSD como en (Al-Mubaid & Chen, 2008) para proporcionar la funcionalidad de texto predictivo y en (Sevens et al., 2016) para la selección de un pictograma correcto.

3 Descripción de la investigación propuesta

La investigación propone como contribución un enfoque de simplificación léxica el cual se ha concretizado en el sistema EASIER que proporciona un soporte sistemático al seguimiento de pautas de accesibilidad cognitiva el cual utiliza recursos de accesibilidad en un dominio genérico y en idioma español.

Tal como muestra la Figura 1, el sistema EASIER incluye dos componentes principales que implementa tareas como la simplificación léxica y la desambiguación del sentido de las palabras. Además, en este trabajo se han creado recursos en el ámbito de la accesibilidad como un diccionario y un corpus los cuales se describirán en esta sección.

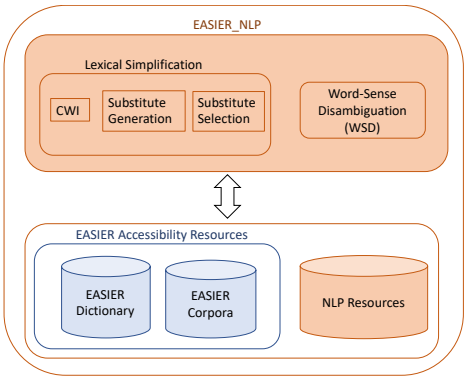

Figura 1: Arquitectura de la investigación propuesta

3.1 Simplificación Léxica

Este módulo de simplificación léxica detecta las palabras complejas en un texto, ofreciendo de estas, sinónimos sencillos. Este sistema sigue el enfoque de Shardlow (Shardlow, 2014) para la simplificación léxica. A continuación, se describen los distintos pasos seguidos.

3.1.1 Identificación de palabras complejas (CWI)

En la tarea de CWI se sigue un enfoque supervisado con un SVM lineal. Dicho algoritmo se ha seleccionado por sus buenos resultados en clasificar en comparación con otros clasificadores. Para entrenar y validar el algoritmo, se ha experimentado principalmente con dos conjuntos de datos: uno de los recursos es el conjunto de datos del workshop BEA 2018 en la tarea CWI. El otro recurso utilizado es un

[1] http://simplext.taln.upf.edu/

[2] http://www.first-asd.eu/

corpus creado en el marco de esta investigación. Este corpus proporciona un conjunto de datos que consta de instancias que proveen una palabra objetivo y su correspondiente clasificación; además cada palabra ofrece sinónimos sencillos según los anotadores. Este recurso es elaborado por expertos en Lectura Fácil y Lenguaje Sencillo.

Siguiendo un enfoque supervisado, cada instancia se representa como un conjunto de características para distinguir entre palabras simples y complejas. Según experimentación hecha hasta el momento en la cual se exploraron distintas propuestas de características (Alarcon et al., 2019), la propuesta final es:

- Característica de Longitud: Longitud de la palabra.
- Característica Booleana: Si una palabra tiene letras mayúsculas.
- Característica de Word2Vec: Vectores de un modelo de 300 dimensiones, entrenado con el Spanish Billion Word Corpus (Cardellino, 2016).
- Característica de BERT: Vectores de un modelo pre-entrenado BERT[3]. Se utiliza primero un modelo multilingüe BERT de 12 capas pre-entrenado antes de extraer los vectores de palabras añadiendo las últimas cuatro capas y utilizando las primeras 480 dimensiones del modelo.
- Además, con el objetivo de incluir una característica que esté basada en recursos de Lectura Fácil, se crea un diccionario denominando. Este diccionario proporciona una característica binaria básica en la que una función verifica si una palabra dada es compleja o no. Este diccionario ha sido creado a partir de fuentes de literatura de Fácil Lectura.

3.1.2 Generación de Sustitutos

La etapa de Generación de Sustitutos se sigue una estrategia de consulta a base de datos lingüísticas, fusionando dos recursos lingüísticos: Babelnet (Navigli & Ponzetto, 2010) y Thesaurus[4]. Al terminar esta etapa se tiene una lista de sinónimos asociada a la palabra compleja objetivo.

3.1.3 Selección de Sustitutos

La etapa de Selección de Sustitutos recibe la lista de sinónimos producida en la etapa anterior. Con esta lista, el objetivo es seleccionar el sinónimo óptimo para reemplazo, teniendo en cuenta factores de simplicidad y teniendo en cuenta el contexto. En relación al contexto se utiliza una métrica de similitud semántica con la ayuda del modelo Word2Vec utilizado en la etapa CWI.

3.2 Desambiguación del sentido de las palabras (WSD)

Con el objetivo de encontrar una definición contextualizada a las palabras complejas detectadas en la etapa de CWI, se proporciona un enfoque para la tarea de realizar desambiguación del sentido de las palabras WSD).

Después de explorar distintos enfoques (Alarcon, Moreno, & Martínez, 2020) se utiliza un modelo pre-entrenado BERT que se utilizó en la etapa de CWI. El pipeline utilizado se describe a continuación. Inicialmente se enmascara la palabra en la oración a la que pertenece, posteriormente el modelo predice qué palabras pueden ocupar el lugar de la palabra enmascarada. Esto devuelve como resultado una lista de palabras que comparten un significado en común, logrando desambiguar la palabra objetivo. Como siguiente paso, se extraen las palabras de la oración que tengan contenido léxico y las agregamos a la lista.

Se utilizan dos diccionarios: el diccionario de la Real Académica de la Lengua (RAE)[5] y el Diccionario Fácil[6], un diccionario de definiciones en Lectura Fácil creado por la asociación Plena inclusión Madrid[7] a través de sus expertos y usuarios con discapacidad cognitiva.

Por último, indicar que este enfoque de simplificación léxica incluyendo además la desambiguación del sentido de las palabras se está concretizado en el sistema EASIER, un sistema desarrollado en escenario web y móvil[8] el cual proporciona una prueba de concepto para probar la idoneidad de la investigación propuesta. El sistema EASIER proporciona ayuda a las personas para comprender mejor los textos, básicamente proporciona simplificación léxica de los textos en español ofreciendo

[3] https://github.com/shehzaadzd/pytorch-pretrained-BERT

[4] www.thesaurus.altervista.org/

[5] https://dle.rae.es/

[6] http://www.diccionariofacil.org/

[7] https://www.plenainclusion.org/

[8] http://163.117.129.208:8080/

distintas ayudas a la comprensión. En la evaluación de este sistema se cuenta con la participación de personas mayores y personas con discapacidad intelectual.

4 Metodología y experimentos propuestos

4.1 Metodología

La metodología que se propone para la consecución de esta tesis se presenta a continuación:

1. Estudio y revisión del estado del arte: Se realiza un estudio de la literatura existente sobre métodos de accesibilidad en el contenido textual y métodos de PLN.

2. Búsqueda, adaptación e integración de recursos:

Búsqueda de recursos en dominio genérico y español para poder realizar un análisis de los métodos propuestos.

Adaptación o creación de recursos para orientar los resultados a utilizar en el ámbito de la accesibilidad cognitiva.

4. Definición y desarrollo de enfoque propuesto: Diseño de una arquitectura modular.

5. Evaluación: Utilización de recursos existentes y recursos generados para llevar a cabo la experimentación. Evaluación del sistema llevando a cabo una comparación de los resultados obtenidos con los ya existentes.

4.2 Experimentos propuestos

4.2.1 Simplificación léxica

En la tarea de CWI, se han definido dos familias de experimentos.

- Comparación de resultados según enfoque propuesto (clasificador y propuesta de características), y los sistemas de la tarea del workshop de BEA 2018. (Alarcon et al., 2019).
- Comparación de resultados entre utilizar el conjunto de datos obtenidos del corpus creado en el marco de la investigación, y el conjunto de datos de BEA workshop.

Siguiendo con la etapa de la Generación de Sustitutos, se realiza una experimentación utilizando el corpus creado. El procedimiento es extraer instancias, cada una conteniendo una oración, una palabra objetivo y tres a cuatro sinónimos sugeridos por un experto lingüista, para finalmente calcular métricas como accuracy, precisión, recall y F1-score.

Por último, para la experimentación de la etapa de la Selección de Sustitutos, se evalua si los candidatos que el sistema seleccione concuerden con los sinónimos que sugiere el gold-standard.

4.2.2 Desambiguación del sentido de las palabras (WSD).

Para la evaluación del módulo de WSD se cuenta con la participación de expertos lingüistas. A partir de conjunto con un número suficiente de oraciones asociadas a su palabra objetivo y definición seleccionada, expertos lingüistas indican si la definición seleccionada es la correcta.

5 Elementos de investigación específicos propuestos para discusión;

Además de la discusión sobre el enfoque, métodos y resultados obtenidos, son varias las cuestiones de investigación que se plantean, entre las que destacan las siguientes:

- ¿Es posible optimizar la simplificación léxica a las personas con discapacidad cognitiva a través de: combinar enfoques de aprendizaje automático con el uso de recursos como el diccionario y corpus creados en el marco del trabajo de investigación?
- ¿Qué impacto puede tener en los resultados del enfoque de la investigación propuesta al realizarse en un dominio específico, en vez de uno genérico?

Agradecimientos

Este trabajo está financiado por el Programa de Investigación del Ministerio de Economía y Competitividad, (Proyecto DeepEMR TIN2017-87548-C2-1-R) y la ayuda “Tecnologías Accesibles” por Indra y Fundación Universia.

Bibliografía

Al-Mubaid, H., & Chen, P. (2008). Application of word prediction and disambiguation to improve text entry for people with physical disabilities (assistive technology). *International Journal of Social and Humanistic Computing*, *1*(1), 10–27.

Alarcon, R., Moreno, L., & Martínez, P. (2020). Word-Sense disambiguation system for text readability. *DSAI 2020 (9th International Conference on Software*

Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion). ACM Digital Library (IN PRESS).

Alarcon, R., Moreno, L., Segura-bedmar, I., & Martínez, P. (2019). Lexical simplification approach using easy-to-read resources. *Procesamiento Del Lenguaje Natural*, 95–102. https://doi.org/10.26342/2019-63-10

Barbu, E., Martín-valdivia, M. T., Martínez-cámara, E., & Ureña-lópez, L. A. (2015). Language technologies applied to document simplification for helpingautistic people. *Expert Systems With Applications*, *42*(12), 5076–5086. https://doi.org/10.1016/j.eswa.2015.02.044

Cao, R., Bai, J., & Shinnou, H. (2019). *Semi-supervised learning for all-words WSD using self-learning and fine-tuning*. (Paclic 33), 356–361.

Cardellino, C. (2016). *Spanish {B}illion {W}ords {C}orpus and {E}mbeddings*. Retrieved from https://crscardellino.github.io/SBWCE/

Carroll, J., Minnen, G., Canning, Y., Devlin, S., & Tait, J. (1998). Practical Simpliication of English Newspaper Text to Assist Aphasic Readers. *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, (June 2013), 7–10. Retrieved from http://alpha.mic.dundee.ac.uk/~slanger/workshop.html

Chen, P., Bowes, C., & Brown, D. (2009). *A Fully Unsupervised Word Sense Disambiguation Method Using Dependency Knowledge*. (June), 28–36.

Cheng, K. (2019). Multilingual Complex Word Identification: Convolutional Neural Networks with Morphological and Linguistic Features. *The 12th International Conference on Recent Advances in Natural Language Processing*, (83–89).

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.

Du, J., Qi, F., & Sun, M. (2019). *Using BERT for Word Sense Disambiguation*. (1).

EU. (2011). How to write clearly. Retrieved from https://op.europa.eu/en/publication-detail/-/publication/c2dab20c-0414-408d-87b5-dd3c6e5dd9a5

Inui, K., Fujita, A., Takahashi, T., Iida, R., & Iwakura, T. (2003). Text Simplification for Reading Assistance. *Proceedings of the 2nd International Workshop on Paraphrasing*, *16*, 9–16. https://doi.org/10.3115/1118984.1118986

Lesk, M. (1986). *Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone*. 24–26.

Medina, J., Saggion, H., Schuurman, I., Sevens, L., O'Flaherty, J. J., De Vliegher, A., & Daems, J. (2016). *Towards Integrating People with Intellectual Disabilities in the Digital World*. (Id), 348–357. https://doi.org/10.3233/978-1-61499-690-3-348

Moradi, B., Ansari, E., & Zabokrtský, Z. (2019). Unsupervised Word Sense Disambiguation Using Word Embeddings. *PROCEEDING OF THE 25TH CONFERENCE OF FRUCT ASSOCIATION*.

Moro, A., & Navigli, R. (2015). *SemEval-2015 Task 13 : Multilingual All-Words Sense Disambiguation and Entity Linking*. (SemEval), 288–297.

Navigli, R., Jurgens, D., & Vannella, D. (2013). *SemEval-2013 Task 12 : Multilingual Word Sense Disambiguation*. *2*(SemEval), 222–231.

Navigli, R., & Ponzetto, S. (2010). BabelNet: Building a very large multilingual semantic network. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, (July), 216–225. Retrieved from http://dl.acm.org/citation.cfm?id=1858704

OCDE. (2013). Resultados del informe PIAAC

de la OCDE. Retrieved from http://www.educacionyfp.gob.es/prensa/actualidad/2013/10/20131008-piaac.html

Paetzold, G. H., & Specia, L. (2015). *Unsupervised Lexical Simplification for Non-Native Speakers*. 3761–3767.

Paetzold, G. H., & Specia, L. (2017). A survey on lexical simplification. *Journal of Artificial Intelligence Research*, *60*, 549–593. https://doi.org/10.1613/jair.5526

Paetzold, G., & Specia, L. (2016). SemEval 2016 Task 11: Complex Word Identification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 560–569. Retrieved from http://aclweb.org/anthology/S16-1085

Pavlick, E., & Callison-Burch, C. (2016). *Simple PPDB : A Paraphrase Database for Simplification*. (In Proceedings of the 54th ACL), 143–148.

Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L., & Drndarevic, B. (2015). Making It Simplext: Implementation and Evaluation of a text Simplification System for Spanish. *ACM Transactions on Accessible Computing*, *6*(4), 1–36. https://doi.org/10.1145/2738046

Sevens, L., Jacobs, G., Vandeghinste, V., Schuurman, I., & Van Eynde, F. (2016). *Improving Text-to-Pictograph Translation Through Word Sense Disambiguation*. 131–135.

Shardlow, M. (2014). A Survey of Automated Text Simplification. *International Journal*, (Special Issue on Natural Language Processing), 58–70. https://doi.org/10.14569/SpecialIssue.2014.040109

UNE. (2018). Asociación Española de Normalización, UNE 153101:2018 (Easy to read. Guidelines and recommendations for the elaboration of documents). Retrieved November 18, 2020, from https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0060036

W3C. (2019). Web Content Accesibility Guidelines (WCAG). Retrieved from https://www.w3.org/WAI/standards-guidelines/wcag/

W3C. (2020). *Grupo de trabajo de accesibilidad para discapacidades cognitivas y de aprendizaje (COGA TF)*. Retrieved from https://www.w3.org/TR/coga-usable/

Yimam, S. M., Stajner, S., Riedl, M., & Biemann, C. (2017). Multilingual and Cross-Lingual Complex Word Identification. *Recent Advances in Natural Language Processing*, 813–822. https://doi.org/10.26615/978-954-452-049-6-104