KEYWEXT: A Multilingual Keyword Extraction Service based on Word Embeddings

KEYWEXT: un Servicio Multilingüe de Extracción de Palabras Clave basado en Word Embeddings

Eva Martínez Garcia, Luis Talegón, Iván Cañaveral, Pablo Martínez, Paul Goldbaum SEEDTAG

c/Marqués de Valdeiglesias 6, 28004, Madrid (Spain) {evamartinez, luis, ivancanaveral, pablomartinez, paul}@seedtag.com

Abstract: Contextual Advertising utilizes the content a user is seeing to understand their interest in real-time to serve relevant advertising. A good representation of this context is the first step to achieve a more precise selection of suitable advertisements that is relevant to the content. We present the description of the SEEDTAG's keywords extractor system demonstration: KEYWEXT. It uses state-of-the-art multilingual BERT-based positional embeddings to help contextualize advertising campaigns by retrieving those n-grams that best represent the content of the document. This leads to more relevant advertising while being respectful with the user.

Keywords: keywords, automatic extraction, word embeddings, multilingual, word2vec, sentence-BERT.

Resumen: La Publicidad Contextual utiliza el contenido que un usuario está viendo para entender su interés en tiempo real. Una buena representación del contexto que un usuario está leyendo en un momento puntual es el primer paso para una mejor selección de los anuncios más adecuados. Presentamos la descripción de la demostración del sistema de extracción de palabras clave de SEEDTAG: KEY-WEXT. Utiliza técnicas de estado del arte sobre embeddings posicionales de modelos multilingües basados en BERT para ayudar a contextualizar campañas publicitarias al extraer aquellos n-gramas que mejor representan el contenido de un documento. De esta manera se asegura una publicidad relevante a la vez que respetuosa con el usuario.

Palabras clave: palabras clave, extracción automática, word embeddings, multilingüe, word2vec, sentence-BERT.

1 Introduction

Contextual Advertising technologies allows brands to reach their target users in the right context by what the user is seeing in that moment. If we can understand that context, we will be able to serve more suitable advertisements thus improving their experience as well as the advertiser impact. A good understanding of the context is key to selecting the most suitable ads for a page and leads to an improved user experience and advertisement impact.

In a digital scenario, the user's context can be a web page showing a news article, a blog post, or an encyclopedia entry. Although this context is nowadays multimodal: text, images, video, etc., especially in the case of professionally-produced content, the text still holds some of the most important part of the information. Natural Language Processing (NLP) techniques can help us to categorize the suitability of a web article taking into account the context of what the user is looking at and without the need to make use of any personal data.

Contextual Advertising strategies typically rely either on well-known taxonomies¹

¹https://www.iab.com/guidelines/ content-taxonomy/

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

or on the targeting of vertical domains (automotive, sports, etc.). The solutions based on these resources are usually rigid and limit the precision that can be achieved are thus insufficient to achieve a higher level of contextualization. We present the description of the SEEDTAG's keywords extraction system demonstration: KEYWEXT. A web service able to extract the most relevant words, bigrams, and trigrams from a web article by using the information from pre-trained word embeddings. These words will help to enhance the context information available for improving the advertising contextualization workflow.

The rest of the paper is structured as follows: Section 2 describes the problem more in detail. Section 3 describes the KEY-WEXT service and its integration in SEED-TAG's contextualization workflow and Section 4 shows some examples of the service functionality. Finally, Section 5 draws conclusions and points out some future work.

2 Motivation

As users, we are used to being surrounded by ads related to our latest search or by the latest sites that we visited. This information does not normally match the website content that we are visiting, and is many times distracting or even annoying.

Since SEEDTAG cares about the privacy of web users, we focus on understanding the context by implementing a cookieless contextualization strategy. We believe that the important information to decide which ad to serve in a particular site comes from the information that the user is seeing at that particular moment. Looking into the text present in a web scenario, we need to understand the information the user is reading in a particular moment to select the most suitable advertisement. Finding the keywords of a text plays an important role in improving the representation of this context. Informally, we understand a set of keywords as the set of phrases that depict the main information from a text. There exist several keywords or keyphrases extraction methods (Campos et al., 2020; Mihalcea and Tarau, 2004). Although these approaches are fast and easy to apply, they are highly language-dependent and many times return noisy lists of words that are difficult to use.

Word embeddings have shown their po-

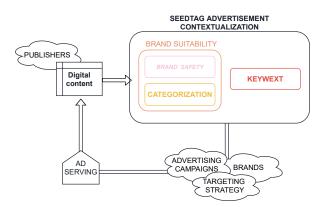


Figure 1: SEEDTAG advertisement contextualization workflow.

tential for solving different tasks (Martínez Garcia et al., 2017; Sun et al., 2019) that they were not trained for. Also, they provide support for a multilingual scenario as well (Reimers and Gurevych, 2020). Following the transfer-learning trend, KEYWEXT uses pre-trained word embeddings to build a context representation vector for a whole document and to retrieve its closest content *n*-grams as the keywords list. Our extractor is not the first one in using BERT-like models (Devlin et al., 2019) to extract keywords (Grootendorst, 2020) but, for the best of our knowledge, KEYWEXT is the first in taking advantage of the positional embeddings and the local contextual information that they provide to retrieve better and more relevant words.

3 System Description

SEEDTAG's contextualization workflow goes as follows. When we have digital content where it is possible to serve advertising, we process the content to assess its suitability and to get all the information needed to select the most adequate advertising campaign.

Figure 1 shows the SEEDTAG's contextualization workflow. The suitability of a particular digital content is measured by analyzing two main features: brand safety and content adequacy. On the one hand, brand safety measures if the content is safe to be related to a brand or product. On the other hand, content adequacy measures how much the content is related to the topics that the brand or product wants to be associated with. Content adequacy is also understood as the categorization of the content. Finally, we use the KEYWEXT service to extract the keywords and keyphrases from the content.

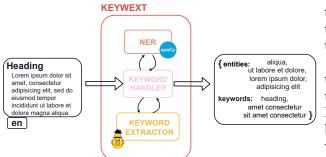


Figure 2: KEYWEXT web service.

Both the categorization and brand safety modules as well as KEYWEXT work directly and in parallel on the text from the digital content. Then, their outputs are combined to feed the contextualization flow. In particular, the context information from KEY-WEXT is used to refine targeting strategies and brand positioning in order to select the best advertisements according to the current context that a user is seeing.

KEYWEXT is a Python² web service built using the Tornado³ web framework. Figure 2 shows the architecture of the system. The service has different actors working together. When it receives a request with a text extracted from a web article and its detected language, the *KeyWord Handler* passes the information to a Spacy⁴ Named Entity Recognizer and the KeyWord Extractor. Then, these modules obtain the list of Named Entities and keywords respectively, that the Key-Word Handler will use to build the response.

3.1 Key Words Extraction

We want to retrieve the most relevant ngrams from a text. Thus, we need to understand the text to select the most suitable content words or n-grams from the text that best represent it.

KEYWEXT performs the keyword extraction in two steps:

- 1. Build a vector representation from the whole input text.
- 2. Retrieve the closest words, bigrams, and trigrams to the text representation.

The first step is done by using a sentence-BERT (Reimers and Gurevych, 2019) pretrained model. KEYWEXT sums the sentence embedding of each sentence in the text to obtain the document embedding.

The second step is done by calculating the distance among the content words from the input text to the document embedding. KEYWEXT uses the BERT positional embeddings for the input tokens to obtain the words, bigrams, and trigrams embeddings and the cosine similarity as distance. These embeddings are the result of summing the embeddings from the tokens that form a particular word, bigram, or trigram. We decided not to consider *n*-grams with n > 3to control the sparseness and the quantity of the possible combinations when checking and calculating distances. Using the positional embeddings from the sentence-BERT models gives the service a local idea of the context of the text. Even though it is not yet a document-level context, this approach allows KEYWEXT to have a broader vision without conflating different senses of a word in the same embedding. In short, that will help to better disambiguate the keyword choice and to produce more adequate results.

3.2 Multilingualism

KEYWEXT is also multilingual. Having a service that is able to handle requests in different languages is crucial for its integration within SEEDTAG's workflow to cover the languages of the countries the company operates in.

Multilingualism is achieved by using a multilingual sentence-BERT-based embedding model (Reimers and Gurevych, 2020) to retrieve the keyword set from articles in different languages.

4 Sample of Keyword Extraction Functionality

We show some examples of the KEYWEXT functionality on some of the most relevant languages for SEEDTAG.

If we process the following short text in English:

How the suspension of the AstraZeneca vaccine is affecting the inoculation drive in each Spanish region. Regional authorities have administered 5.7 million doses and fully vaccinated nearly 1.7 million people, but the jabs for essential workers have been put on hold due to the decision to halt the use of the Anglo-Swedish medication.[...]

²https://www.python.org/

³https://www.tornadoweb.org/

⁴https://spacy.io/

The KEYWEXT service returns the following keyword list:

astrazeneca, administered, suspension, vaccine, astrazeneca vaccine, authorities have administered

Notice how a generic open-domain pretrained word embedding model can detect a recent Named Entity like *astrazeneca* as a relevant element of the text. If a different kind of embedding model such as *word2vec* (Mikolov et al., 2013) had been used, this adaptation would not have been possible due to vocabulary coverage restrictions.

Moving to Spanish texts, when processing a negative news piece about an attack in Burkina Faso discussing the death of two journalists :

Dos periodistas españoles mueren asesinados en un ataque en Burkina Faso. Un grupo de hombres armados asaltó el convoy de los reporteros David Beriain y Roberto Fraile en dos camionetas y una decena de motos.[...]

We obtain the following set of keywords using our KEYWEXT service:

ataque, asesinados, mueren, periodistas, viajaban, asaltó, españoles mueren, periodistas españoles mueren, ataque en Burkina

Although these words or phrases can seem trivial, once fed into our contextualization models they reinforce their knowledge about potential harmful content and allow SEED-TAG to help advertisers better design their Contextual Advertising strategies.

5 Conclusions and Future Work

We presented KEYWEXT, a keywords extraction system that takes advantage of pretrained word embeddings to retrieve the most relevant *n*-grams from an article. These extracted keywords feed into SEEDTAG's contextual advertising workflow to identify the most suitable matches among brands, their advertising campaigns and web articles.

KEYWEXT is a web service that uses sentence-BERT-based pre-trained models to understand the context of an article beyond the sentence level and, then, retrieve the closest words, bigrams, and trigrams of the document. Also, KEYWEXT takes advantage of the multilingual sentence-BERT models to handle articles in different languages.

New versions of KEYWEXT will improve handling document-level information: using document-level text representations, taking into account topic fluctuations when producing the set of the top keywords, etc. The new features will improve the extraction of keywords for longer and more complex texts.

References

- Campos, R., V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.
- Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NACL2019*.
- Grootendorst, M. 2020. Keybert: Minimal keyword extraction with bert.
- Martínez Garcia, E., C. Creus, C. España-Bonet, and L. Màrquez. 2017. Using word embeddings to enforce document-level lexical consistency in machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108.
- Mihalcea, R. and P. Tarau. 2004. Textrank: Bringing order into text. In *Proceedings* of *EMNLP2004*.
- Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 26: 27th NIPS.
- Reimers, N. and I. Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings* of *EMNLP2019*.
- Reimers, N. and I. Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of EMNLP2020*.
- Sun, C., X. Qiu, Y. Xu, and X. Huang. 2019. How to fine-tune bert for text classification? In China National Conference on Chinese Computational Linguistics.