ALEXSIS: A Dataset for Lexical Simplification in Spanish

Daniel Ferrés[1], **Horacio Saggion**[1]

[1]Large Scale Text Understanding Systems Lab @ TALN-DTIC - Universitat Pompeu Fabra, Barcelona (Spain)

Introduction

1. **Lexical Simplification (LS)**
   LS is the task of replacing difficult words with easier to read expressions while keeping meaning.
   LS Tasks in execution order:
   1. Complex Word Identification (CWI) - (not used)
   2. **Substitution Generation (SG)**
   3. **Substitution Selection (SS)**
   4. **Substitution Ranking (SR)**
   5. Morphological Generation and Context Adaptation (not used)
2. **Dataset Compilation**
   ALEXSIS has 381 instances with: a sentence, a target complex word and 25 annotated substitutions
3. **Evaluation of LS Approaches**
   - Combinations of SG, SS, and SR: 1) SG, 2) SG+SS, 3) SG+SR, and 4) SG+SS+SR
   - Datasets: ALEXSIS and EASIER [Alarcón et al., 2021]

Compilation Procedure

1. **<sentence, complex word> pair extraction**
   588 pairs were extracted from the Spanish CWI 2018 Shared Task dataset [Yimam et al., 2018]. The complex word had to be a single word marked by 5 or more native language annotators.
2. **Manual judgment process**
   2 computational linguists experts decided if each complex word in the context was: simplifiable, not simplifiable or dubious. The process resulted on a set of 402 simplifiable pairs that was then reduced after some additional filtering.
3. **Set the Annotation Task**
   The annotation task consists to propose a single word that is a valid simpler synonym or replacement for the complex word in the context but is easier to understand. If a single-word is not possible then phrases or multi-words are allowed. Otherwise the complex word should be returned.
4. **Pilot Study**
   It was used to estimate the average time to complete the task: about 45 seconds per instance
5. **Final Study**
   The *prolific.co* platform was used to hire annotators with these requirements: Spanish native speakers and with an undergraduate or a graduate degree as the minimum education achieved. 25 annotators per sentence were hired.
6. **Dataset Statistics**
   - 381 instances in the dataset
   - 356 unique target words: 333 (once), 21 (twice), 2 (3 times)
   - 9,524 substitutions (and 3,918 of them are unique).
   - 10.28 average number of unique synonyms per instance

ALEXSIS instance example

sentence	*Sufrió una importante reducción en su capacidad para poder* ***acogerse*** *a las normas de la FIFA para los estadios de fútbol.*
complex word	***acogerse***
annotations	adaptarse (6), refugiarse (2), apegarse (2), ampararse (2), aceptar (2), incorporarse (2), sumarse, recurrir, obedecer, cumplir_con, asimilarse, aplicarse, amparar, admitirse, aceptarse

Substitution Generation

1. **Thesaurus-based system**. [Ferrés et al., 2017].
2. **LSBert-es**
   It is an Adaptation of LSBert [Qiang et al., 2020] for Spanish with BETO and other resources.
3. **Single Transformers**
   This approach is based on LSBert [Qiang et al., 2020]. It uses the Masked Language Model (MLM): concatenates the original sentence with the same sentence with the complex word. The Spanish pre-trained models used were: BETO,mBERT, SpanBERTa , BERTIN, RoBERTa-base-BNE (RbaseBNE) RoBERTa-large-BNE (RlargeBNE)
4. **Combination of Results of Transformers**: combines them with *Union* (∪) or *Intersection* (∩).

Substitution Selection

1. **Morphological filtering**:
   It uses the Freeling's morphological data to: 1) filter out morphological variations of the complex word and 2) combine all the morphological variations of candidates.
2. **POS-tag filtering**
   This approach applies POS-tagging to the original sentence and to new sentences with the candidates instead of the complex word. Then discards the candidates without the same lexical category of the complex word.

Substitution Ranking

1. **Corpus-based Ranking**
   - 3 lists of word frequencies used: SUBTLEX-ESP, ESWIKI-2014, OpenSubtitles-2016
2. **BERT fine-tuned (BETO)**
   Train./dev./test. with 9,607 words from CWI 2018 dataset. The approach was based on [Bani Yaseen et al., 2021] but using only tokens. It predicts a real number that represents the complexity of the word.

SG Results

System	Potential	Precision	Recall	F1
Baselines [Alarcón et al., 2021]				
Word2vec	0.358	0.019	0.188	0.034
FastText	0.464	0.029	0.289	0.053
Sense2Vec	0.506	0.056	0.298	0.095
BETO	0.348	0.03	0.282	0.054
Our Approaches				
Thesaurus-based	0.198	**0.124**	0.089	**0.104**
LSBert-es (BETO)	0.764	0.027	0.464	0.051
BETO	0.697	0.025	0.422	0.048
SpanBERTa	0.848	0.035	0.601	0.067
mBERT	0.328	0.010	0.161	0.019
BERTIN	0.800	0.033	0.564	0.063
RbaseBNE	0.826	0.035	0.589	0.067
RlargeBNE	0.824	0.035	0.585	0.067
BETO ∪ SpanBERTa	0.782	0.031	0.533	0.059
BETO ∩ SpanBERTa	0.674	0.052	0.406	0.093
SpanBERTa ∪ RbaseBNE	**0.866**	0.036	**0.618**	0.069
SpanBERTa ∩ RbaseBNE	0.808	0.049	0.561	0.090

Table 1:Evaluation on EASIER-500 (top-k=50)

System	Potential	Precision	Recall	F1
Thesaurus-based	0.146	**0.132**	0.021	0.037
LSBert-es (BETO)	0.860	0.047	0.245	0.079
mBERT	0.545	0.023	0.118	0.039
BETO	0.782	0.042	0.213	0.071
SpanBERTa	0.892	0.059	0.308	0.099
BERTIN	0.853	0.057	0.294	0.096
RbaseBNE	0.913	0.061	0.317	0.103
RlargeBNE	0.910	0.061	0.318	0.103
BETO ∪ SpanBERTa	0.876	0.053	0.277	0.089
BETO ∩ SpanBERTa	0.745	0.093	0.188	0.124
BERTIN ∪ RbaseBNE	**0.921**	0.062	0.325	0.105
BERTIN ∩ RbaseBNE	0.837	0.085	0.263	0.129
SpanBERTa ∪ RbaseBNE	0.918	0.064	**0.332**	0.107
SpanBERTa ∩ RbaseBNE	0.881	0.087	0.284	**0.133**

Table 2:Evaluation on ALEXSIS (top-k=50)

Full Pipeline Results

System	Precision	Accuracy	Changed
Thesaurus-based	**0.776**	0.096	0.320
LSBert-es (BETO)	0.236	0.228	0.992
Transformers			
RlargeBNE	0.33	0.312	0.982
SpanBERTa	0.281	0.281	1.0
BERTIN ∩ SpanBERTa	0.338	0.302	0.964
BERTIN ∪ SpanBERTa	0.290	0.290	1.0
Transformers+filterMorpho+filterPOS			
RlargeBNE	0.35	0.326	0.976
SpanBERTa	0.352	0.347	0.996
BERTIN ∩ SpanBERTa	0.402	0.347	0.946
BERTIN ∪ SpanBERTa	0.354	**0.350**	0.996

Table 3:Evaluation on EASIER-500 (top-k=1).

System	Precision	Accuracy	Change
Thesaurus-based	**0.889**	0.089	0.199
LSBert-es (BETO)	0.278	0.278	1.0
Transformers			
RbaseBNE	0.438	0.438	1.0
SpanBERTa	0.409	0.409	1.0
SpanBERTa ∩ RbaseBNE	0.456	0.456	1.0
SpanBERTa ∪ RbaseBNE	0.454	0.454	1.0
Transformers + filterMorpho+filterPOS			
RbaseBNE	0.454	0.451	0.997
SpanBERTa	0.448	0.448	1.0
SpanBERTa ∩ RbaseBNE	0.475	0.461	0.986
SpanBERTa ∪ RbaseBNE	0.469	**0.469**	1.0

Table 4:Evaluation on ALEXSIS (top-k=1).

Conclusions

1. **ALEXSIS Dataset**
   ALEXSIS includes information potentially useful for Lexical Simplicity Ranking but might needs further research with linguistic experts. ALEXSIS has higher number of average unique synonyms per instance compared with EASIER/EASIER-500 (for Spanish) and most of the other datasets in other languages.
2. **Evaluation of LS Approaches**
   Some approaches with transformers for SG evaluated with EASIER-500 showed state-of-the-art results [Alarcón et al., 2021]. Some SG approaches with combination of transformers obtained the best results in the SG evaluation with both ALEXSIS and EASIER-500. The SS approaches applied in combination with SG achieved the best results in the full pipeline evaluation. In most of the experiments the SR approaches do not achieve improvements of the results.

Further Work

1. **Create a new version of the dataset:**
   1) without incorrect substitutions and dubious words, and 2) with new synonyms compiled manually with dictionaries/thesaurus and from results of the transformers-based approaches.
2. **Improve the LS approaches:**
   1) with new or existing state-of-the art algorithms, and 2) with better filtering out wrong candidates in SS tasks.

Acknowledgements

We acknowledge support from the project Context-aware Multilingual Text Simplification (ConMuTeS) PID2019-109066GB-I00/AEI/10.13039/501100011033 awarded by Ministerio de Ciencia, Innovación y Universidades (MCIU) and by Agencia Estatal de Investigación (AEI) of Spain.

Contact Information

- Web: taln.upf.edu
- Emails: {daniel.ferres,horacio.saggion}@upf.edu