UAlberta at LSCDiscovery: Lexical Semantic Change Detection via Word Sense Disambiguation

Daniela Teodorescu, Spencer von der Ohe, Grzegorz Kondrak
Alberta Machine Intelligence Institute, Department of Computing Science
University of Alberta, Edmonton, Canada
{dteodore,vonderoh,gkondrak@ualberta.ca

Abstract

We describe our two systems for the shared task on Lexical Semantic Change Discovery in Spanish. For binary change detection, we frame the task as a word sense disambiguation (WSD) problem. We derive sense frequency distributions for target words in both old and modern corpora. We assume that the word semantics have changed if a sense is observed in only one of the two corpora, or the relative change for any sense exceeds a tuned threshold. For graded change discovery, we follow the design of CIRCE (Pömsl and Lyapin, 2020) by combining both static and contextual embeddings. For contextual embeddings, we use XLM-RoBERTa instead of BERT, and train the model to predict a masked token instead of the time period. Our language-independent methods achieve results that are close to the best-performing systems in the shared task.

1 Introduction

Lexical semantic change discovery is a task with growing interest and applications in various areas, such as natural language processing and lexicography (Schlechtweg et al., 2020). The shared task on semantic change discovery and detection in Spanish (LSCDiscovery) consists of two phases: 1) graded change discovery, and 2) binary change detection (Zamora-Reina et al., 2022). We adopt different approaches for both phases.

The two sub-tasks consider different aspects of lexical semantic change (LSC). The definition for graded change discovery follows Kurtyigit et al. (2021): *given a diachronic corpus pair C1 and C2, rank the intersection of their (content-word) vocabularies according to their degree of change between C1 and C2.* For binary LSC detection, the definition is the same as used in the SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020): *given a target word *w* and two sets of its usages U1 and U2, decide whether *w* lost or gained senses from U1 to U2, or not.* The direction of change is not important in either task. The inputs to the tasks consist of a list of target words and a pair of corpora from different time periods, annotated for the semantic relationship between word usages. The gold labels on a set of target words are inferred from sense frequency distributions derived by clustering the manual annotations (Zamora-Reina et al., 2022). The output in phase 1 is a list of the target words ranked by the amount of change. The output in phase 2 is a list of binary change detection labels per word.

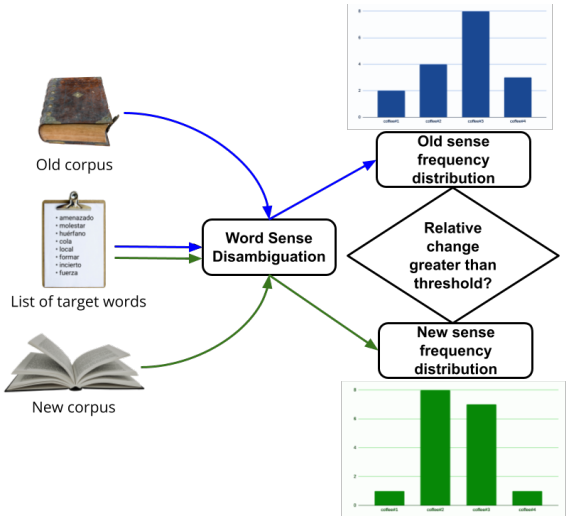

Figure 1: Lexical semantic change detection via WSD.

For graded change discovery, our approach is similar to CIRCE (Pömsl and Lyapin, 2020), the top performing system in the SemEval 2020 task for graded change. We use embeddings and Euclidean distance to obtain rankings. However, we obtain contextual embeddings from token prediction instead of time period prediction. In addition, we use XLM-RoBERTa instead of BERT, because it performs well across a variety of tasks in Spanish (Conneau et al., 2020), and models based on this architecture produce effective contextual embeddings (Ethayarajh, 2019).

Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 180 - 186

For binary change detection, we propose a novel approach based on framing LSC discovery as word sense disambiguation (WSD) problem, which is the task of determining the meaning of words in context given a sense inventory (Navigli, 2009). Using a recently-proposed WSD system, AMuSE (Orlando et al., 2021), we identify the sense of each target word in context to determine if senses were lost or gained over time. Following the theory of Hauer and Kondrak (2020), we posit that wordnet-type sense inventories match the intuitions of the annotators of the shared task data. Our approach has the advantage of being interpretable, providing interesting insights into the nature of lexical semantic change by identifying specific senses that appear or disappear in texts over time.

Our systems are highly competitive. For phase 1, our system achieves 0.5731 correlation between the ranked words and ground truth on the test set, which would put it in third place, based on our own evaluation performed after the submission deadline. For phase 2, our system obtains F-score of 88% on the development set, and 71% on the evaluation set, which ranks it as second according to the main metric.

2 Related Work

In the SemEval 2020 task for graded change discovery, the CIRCE system performed the best (Schlechtweg et al., 2020). The system ensembles static and contextual embeddings. Static embeddings with Skip-Gram with Negative Sampling (Mikolov et al., 2013) are obtained for each corpus. These embeddings are then aligned using Orthogonal Procrustes analysis (Schönemann, 1966), and the Euclidean distance is found between aligned embeddings. Contextual embeddings from the masked language model BERT (Devlin et al., 2019) are used to classify the time period of sentences, as time specific features are useful to learn. Then, embeddings are extracted from the last hidden layer for each target word. To obtain a distance, the Euclidean distance is computed pairwise between the embeddings from the two corpora and the distances are averaged. The target words are then ranked for both types of embeddings. Finally, the rankings are combined by a weighted average to obtain the final ranking.

We modify the CIRCE approach for our graded change discovery system by using XLM-RoBERTa (Conneau et al., 2020) to obtain contextual embeddings. XLM-RoBERTa is a multilingual masked language model. It uses the same bidirectional transformer architecture as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), and has the same number of layers and size of layers as RoBERTa. However, compared to BERT and RoBERTa it has a larger vocabulary of 250,000 tokens, and employs the SentencePiece tokenizer (Kudo and Richardson, 2018). Additionally, it is trained on 100 languages, instead of just English (Conneau et al., 2020).

Systems for binary change detection commonly use embeddings for semantic representations, with type embeddings often outperforming token embeddings (Schlechtweg et al., 2020). Some approaches ensemble models (Martinc et al., 2020; Pömsl and Lyapin, 2020) or use a topic model (Sarsfield and Tayyar Madabushi, 2020). Nulty and Lillis (2020) detect change by considering the relationship between nodes in a semantic network graph. Orthogonal Procrustes analysis (Schönemann, 1966) and vector initialization (Kim et al., 2014) are techniques that can be used to align the semantic representations. As a distance metric between embeddings, cosine and Euclidean distance are commonly used.

Our approach based on applying WSD for binary change detection is novel. A previous work has performed word epoch disambiguation for determining changes in word usages overtime, but this task predicts the time period (epoch) for instances (Mihalcea and Nastase, 2012). Some previous work considers changes of senses overtime; however, rather than WSD, they apply sense induction (Mitra et al., 2014; Tahmasebi and Risse, 2017), topic modelling (Lau et al., 2012), or Bayesian models (Frermann and Lapata, 2016).

3 Methods

In this section, we describe our methods separately for each of the two phases. Our code is available for public use.[1]

3.1 Phase 1: Graded Change

We follow the approach implemented in CIRCE (Pömsl and Lyapin, 2020), which has been shown to perform well across a number of languages. We use both static and contextual embeddings because a combination of rankings from both embeddings

[1] https://github.com/sazzy4o/ualberta-lscdiscovery

outperforms the ranking from either. We rank target words based on the distance between embeddings of both corpora.

To obtain static embedding rankings, we use the same methods as CIRCE for the following 3 steps. First, we train static embeddings using Skip-Gram with Negative Sampling (Mikolov et al., 2013) for the lemmatized version of each corpus, and align embeddings from both corpora. Second, we obtain the Euclidean distance between the aligned embeddings. Finally, we rank the target words by the Euclidean distance.

The contextual embeddings used in CIRCE perform poorly compared to the static embeddings. We posit that this is because the time period prediction task is not well aligned with predicting the meaning of words. To address this, we first train a XLM-RoBERTa (Conneau et al., 2020) model to predict randomly masked words in the combined corpus from both time periods. Second, we mask out each instance of a target word, and use our trained model to predict the masked word. From this prediction, we extract the embedding from the features corresponding to the masked token in the last hidden layer outputs of the model. Third, we compute the pairwise Euclidean distance between each pair of target word instances from different time periods. Finally, we rank the target words by the mean distance for each target word.

We follow a similar procedure to CIRCE to obtain the final ranking by ordering the target words by a weighted combination of the static and contextual rankings. However, instead of calculating the weighting based on the accuracy of our contextual model, we tune the weighting to maximize Spearman's rank-order correlation on the development set.

This approach is quite computationally expensive since it requires training the XLM-RoBERTa model. The model takes approximately 3-4 hours using an NVIDIA GeForce RTX 3090 to train. Additionally, it takes approximately 45 minutes to obtain the embeddings for the 60 target words in the test set. The static embeddings are significantly faster to obtain, taking only 3 minutes with an Intel Xeon W-2255 CPU.

3.2 Phase 2: Binary Change

We approach binary semantic change detection as WSD. We implement our approach using AMuSE, a user-friendly end-to-end neural WSD system of Orlando et al. (2021), which incorporates the preprocessing steps of tokenization, lemmatization, and parts of speech (POS) tagging. AMuSE is trained on manual annotations involving English WordNet senses, but thanks to its use of the multilingual XLM-RoBERTa embeddings, it is also applicable to other languages that are represented in BabelNet. According to the Universality Principle of Hauer and Kondrak (2020), there is a one-to-one correspondence between concepts in different languages. We apply AMuSE via the REST API[2] to all sentences that contain the target words.

We depict our process in Figure 1. For each word, we compute its sense frequency distributions in both the old and modern corpora based on the output of the WSD system. If a sense is found in the modern corpus but is missing in the old corpus (or vice versa), a change is deemed to have occurred (label 1). Otherwise, a word has the same set of senses identified in both the old and modern corpus. For each sense, we compute the relative probability change (p_r) as the ratio between the absolute probability difference, and the larger of the two probabilities (Formula 1). The probability of a sense for a target word from the new and old corpora is denoted as p_1 and p_2, respectively.

$$p_r = \frac{|p_1 - p_2|}{\max(p_1, p_2)} \quad (1)$$

The resulting value is compared to a threshold, which we tune on the development data by maximizing F-score. A relative change greater than the threshold (set at 0.65) for any of the word senses indicates that a change occurred for the given word (label 1). Otherwise, we conclude that there is no change (label 0).

The definition of binary change detection suggests that it may be sufficient to determine if the set of senses for a target word remains the same from the old to the modern corpus. We implemented this approach after the submission deadline, and obtained 78% F-score on the development set, which is below the F-score of 88% obtained with our principal method described above.

Additionally, we computed two other metrics for phase 2 after the submission deadline: sense gain and sense loss detection. First, the same methodology is applied for detecting change, as sense gain/loss is only applicable when there is change.

[2] https://nlp.uniroma1.it/amuse-wsd/api-documentation

If change is due to the threshold, we compare if the old or modern probability is greater for sense loss/gain. In scenarios where a sense was missing in either the old sense set or the new sense set, we use the direction of change to detect sense gain/loss. Otherwise, if none of the above scenarios apply, we conclude that there is no sense gain/loss. A lemma can be labelled as having both sense gain and loss. Our approach allows for this by calculating the labels separately, and searching through the senses for a word until gain/loss is detected before assigning the no change label.

We consider our approach for phase 2 as lightweight. Although AMuSE uses XLM-RoBERTa embeddings, we did not have to train them. Approaches that rely on contextual embeddings, such as BERT, may be computationally too expensive to run on all instances (Kurtyigit et al., 2021). Given a few hundred sentences per target word in either corpus, we can run AMuSE on all instances, rather than just a sample. We did not use GPU, and simply ran the script on an Intel Xeon CPU E5-2650 v4. The run time for WSD was approximately a few hours for the development set (20 words) and close to a day for the evaluation set (60 words). WSD results were only computed once and then stored.

Further, we highlight how our approaches for both phases are multilingual. Static embeddings can be trained on the given corpora, and XLM-RoBERTa is multilingual by nature. In phase 2, the AMuSE WSD system allows for state-of-the-art neural WSD in 40 languages. This demonstrates that challenges described by Tahmasebi et al. (2021), such as having a translated corpus to train WSD systems, may not always be the case.

4 Evaluation

We test our methods on the development set, and report the results on the evaluation set. Some results were obtained after the submission deadline.

4.1 Phase 1: Graded Change

We use the tokenized and lemmatized versions of the corpora provided in the competition to obtain contextual and static embeddings, respectively. We use CIRCE's implementation[3] for static embeddings, as well as for combining the predictions between models. We use the implementation of XLM-RoBERTa from the Hugging Face transformers library (Wolf et al., 2020)[4] for contextual embeddings. We initialize the weights of the model to the *xlm-roberta-large* available with the transformers library.[5]

For evaluation, we use Spearman's rank-order correlation coefficient (Bolboaca and Jäntschi, 2006) between our ranking and the provided gold ranking. After tuning weights on the development set, the results of our system are 0.8375 and 0.5731 on the development and evaluation set, respectively. Our results are much better than CIRCE, which achieves a correlation of only 0.1894 averaged over three runs on the evaluation set. Only two submissions to the shared task achieved a higher correlation on the evaluation set.

After analysing the rankings in the development set, we find that *aguantar* and *descendiente* are incorrectly ranked by 7 and 8 positions respectively. Both of these words have a relatively low frequency in the modern corpus. In addition, *descendiente* occurs in the old corpus both as a noun and as a verb. All of the other words are within 6 positions of their correct rank with the majority being 2 or fewer positions from their correct rankings.

	Dev			Eval		
	P	R	F1	P	R	F1
Change	79	100	88	55	100	71
Sense Gain	71	100	83	33	93	49
Sense Loss	30	100	46	50	92	65

Table 1: Results for the binary change tasks (in %) on the development and evaluation sets.

4.2 Phase 2: Binary Change

The results of our method are shown in Table 1. According to the official results, the F-score of 71% on the evaluation set, which is the main metric for binary change detection, ranks our system as second in the competition. It is interesting to note that our approach obtains 100% recall, whereas the baseline provided by the organizers obtains 100% precision on the development set, so whenever the two models agree, their classification is correct. For the optional tasks of sense gain/loss detection, we calculated our results after the official submission deadline. At the time of writing, our results for

[3] https://github.com/mpoemsl/circe

[4] https://huggingface.co/docs/transformers/model_doc/xlm-roberta

[5] https://huggingface.co/xlm-roberta-large

sense gain and loss detection would place third and second, respectively.

According to our error analysis on the development set, our system disagrees with the gold annotation by identifying semantic change in the following three words: *descendiente*, *músculo*, and *reforma*. In each of these cases, new senses were found by AMuSE in the modern corpus; in addition, two senses appear to have been lost for *reforma*. Some instances could be interpreted as genuine lost senses. For example, one of the senses of the noun *reforma*, defined in WordNet 3.0 (Miller, 1995) as "rescuing from error and returning to a rightful course" occurs in the old corpus in the following context: *no puedo enseñar a las niñas más que dos cosas: la* **reforma** *de letra y la fábula mitológica*. This suggests that WSD could be an effective approach for identifying changes in sense inventories.

Further inspection reveals that some instances of a spurious new sense identification may have been caused by incorrect POS tags assigned by AMuSE. The gold annotations seem to consistently assign a single POS tag to each target word. We experimented with a modified approach to binary change detection, which only considers the occurrences in which the assigned POS tag matches the most likely tag for a given lemma in BabelNet (Navigli and Ponzetto, 2010), but the results were slightly lower than for our main method.

5 Conclusion

We presented systems for both graded and binary change discovery in the context of the shared task on Lexical Semantic Change Discovery in Spanish. For the former, we proposed a system based on CIRCE, with the modification of tuning the weights between the static and contextual embeddings, and training the model to predict a masked token rather than the time period. For the latter, we demonstrated that a WSD system can be effective in detecting word meaning changes. Future work could include combining rankings from more than two different models for graded LSC discovery. We would also like to investigate if either of our two methods could be applied to the other of the two subtasks.

Acknowledgements

This research was supported by the Alberta Machine Intelligence Institute (Amii), the Natural Sciences and Engineering Research Council of Canada (NSERC), the Social Sciences and Humanities Research Council (SSHRC), and Alberta Innovates.

References

Sorana-Daniela Bolboaca and Lorentz Jäntschi. 2006. Pearson versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds. *Leonardo Journal of Sciences*, 5(9):179–200.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Lea Frermann and Mirella Lapata. 2016. A Bayesian Model of Diachronic Meaning Change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Bradley Hauer and Grzegorz Kondrak. 2020. Synonymy = translational equivalence. *CoRR*, abs/2004.13886.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Lexical Semantic Change Discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6985–6998, Online. Association for Computational Linguistics.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word Sense Induction for Novel Sense Detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601, Avignon, France. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. Discovery Team at SemEval-2020 Task 1: Context-sensitive Embeddings Not Always Better than Static for Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 67–73, Barcelona (online). International Committee for Computational Linguistics.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 259–263, Jeju Island, Korea. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1020–1029, Baltimore, Maryland. Association for Computational Linguistics.

Roberto Navigli. 2009. Word Sense Disambiguation: A Survey. *ACM Comput. Surv.*, 41(2).

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Paul Nulty and David Lillis. 2020. The UCD-Net System at SemEval-2020 Task 1: Temporal Referencing with Semantic Network Distances. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 119–125, Barcelona (online). International Committee for Computational Linguistics.

Riccardo Orlando, Simone Conia, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli. 2021. AMuSE-WSD: An All-in-one Multilingual System for Easy Word Sense Disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 298–307, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin Pömsl and Roman Lyapin. 2020. CIRCE at SemEval-2020 Task 1: Ensembling Context-Free and Context-Dependent Word Representations. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 180–186, Barcelona (online). International Committee for Computational Linguistics.

Eleri Sarsfield and Harish Tayyar Madabushi. 2020. UoB at SemEval-2020 Task 1: Automatic Identification of Novel Word Senses. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 239–245, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors. 2021. *Computational approaches to semantic change*. Number 6 in Language Variation. Language Science Press, Berlin.

Nina Tahmasebi and Thomas Risse. 2017. Finding Individual Word Sense Changes and their Delay in Appearance. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 741–749, Varna, Bulgaria. INCOMA Ltd.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System*

Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. Lscdiscovery: A shared task on semantic change discovery and detection in spanish. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics.