# Edinburgh at SemEval-2022 Task 1: Jointly Fishing for Word Embeddings and Definitions


**Pinzhen Chen  Zheng Zhao**

School of Informatics, University of Edinburgh  {pinzhen.chen, zheng.zhao}@ed.ac.uk


## SemEval 2022 Task 1

Mickus et al. (2022) organised two reciprocal sub-tasks :

1. **reverse dictionary:** generate a word embedding from a textual definition;
2. **definition modelling:** generate a textual definition from a word embedding.

The task covers **5 languages**: English, Spanish, French, Italian, and Russian, and provides words in **3 embedding architectures**: *sgns* (skip-gram with negative sampling), *char* (character-level autoencoding), and *electra*.

In reverse dictionary, embedding similarity is measured by MSE, cosine and ranking scores. For definition modelling, sense-BLEU, lemma-BLEU and MoverScore are used. Organisers provided a baseline which is a single Transformer block.

## Our *Unified* Approach

A definition and the corresponding word embedding has the **same meaning, presented in different forms** (human words and vectors). We utilize a model we proposed earlier as illustrated in Figure 1. It learns to encode definitions and words **into a shared space**, then generates both forms in a **multi-task** fashion (Chen and Zhao, 2022).

As the workflow indicates, only half of the model is utilized during inference for either sub-task. The system is fairly comparable with the baseline.

Besides the architecture selection, we also create ensembles for reverse dictionary and construct naive n-grams for definition modelling.

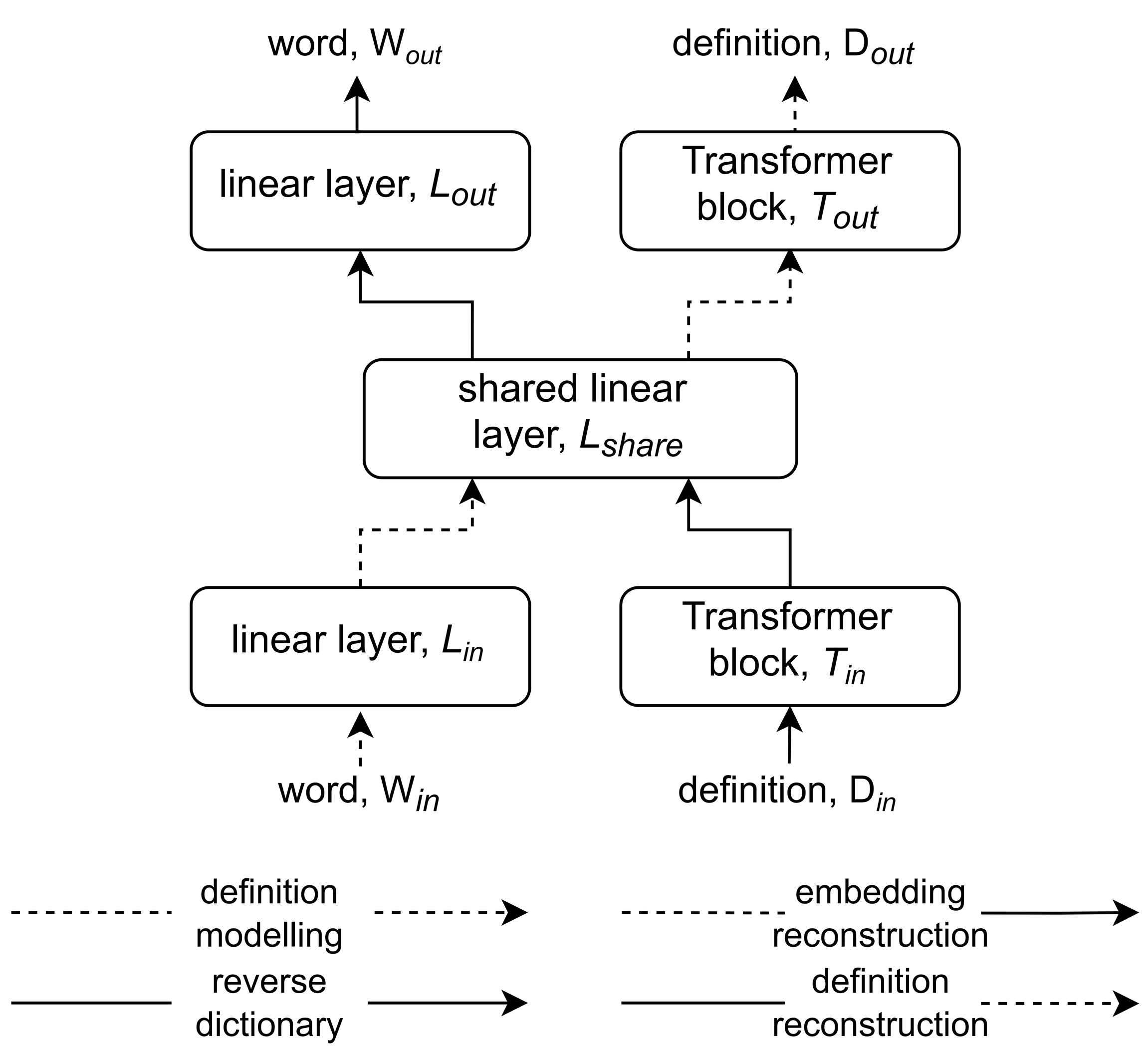



Figure 1. The unified model.

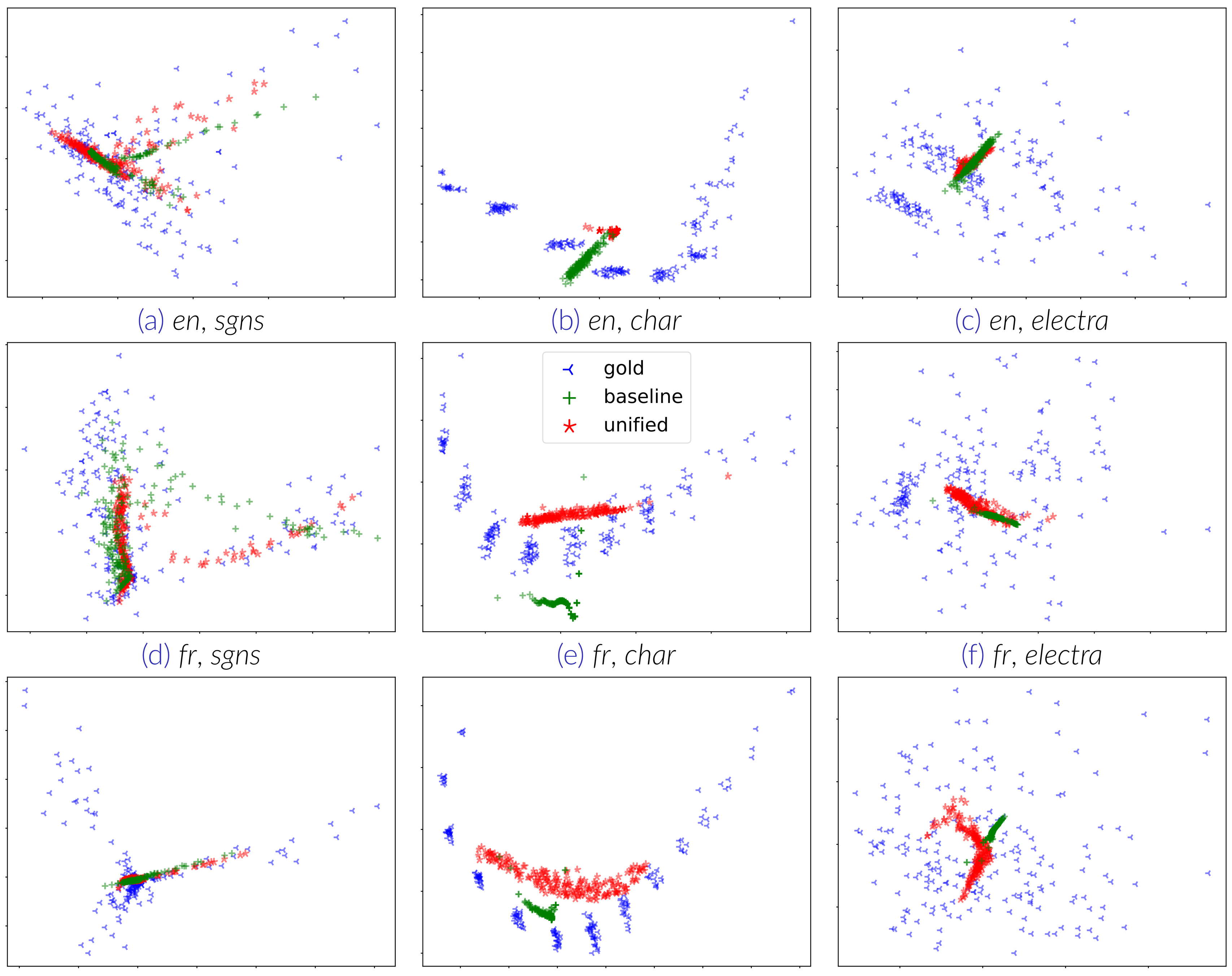
## Results: 8× 🥇, 4× 🥈, 4× 🥉

We submitted to all 18 sub-tracks, and attained first place in 8, second place in 4, and third place in 4. In terms of team ranking, our systems did **better in reverse dictionary** than in definition modelling. Especially, our submissions dominated reverse dictionary with char and electra embeddings.

However, when comparing reverse dictionary's retrieval ranking scores across different embeddings **internally, *sgns* works the best with our models**.

The overall definition modelling performance in this task is poor; our nonsensical n-grams top the French sense-BLEU leaderboard. **BLEU scores are inflated** by stop words and smoothing.

## Visualization of Output Embeddings




(a) *en, sgns* (b) *en, char* (c) *en, electra* (d) *fr, sgns* (e) *fr, char* (f) *fr, electra* (g) *ru, sgns* (h) *ru, char* (i) *ru, electra*

Figure 2. Visualization of gold and output embeddings after 2D PCA, across different languages and embedding architectures.

## Results by Data Features

We inspect our unified model's results categorised by linguistic features: polysemy, part-of-speech, word length, definition length, frequency. Our observation is that **the highest scores from the two sub-tasks emerge in differing categories**, regardless of the feature. We present three features below; more can be found in our paper.

| Polysemy | sgns | | char | | electra | |
|---|---|---|---|---|---|---|
| | cosine | lemma-BLEU | cosine | lemma-BLEU | cosine | lemma-BLEU |
| Yes | 0.232 | **4.34** | 0.804 | **3.20** | 0.836 | **3.61** |
| No | **0.360** | 2.82 | **0.813** | 2.53 | **0.845** | 3.09 |

Table 1. Performance across polysemy annotations for *en*.

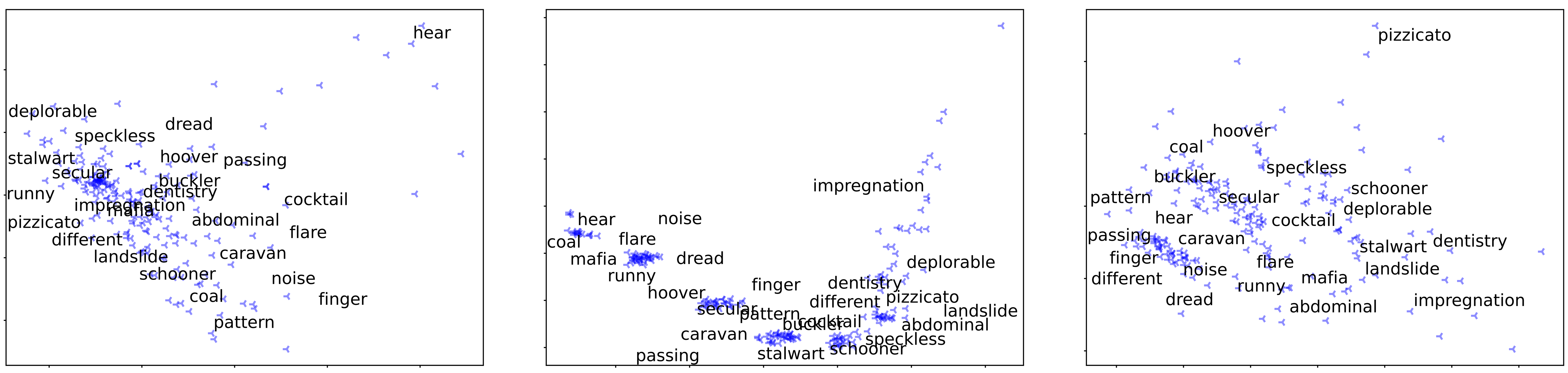
| POS | sgns | | char | | electra | |
|---|---|---|---|---|---|---|
| | cosine | lemma-BLEU | cosine | lemma-BLEU | cosine | lemma-BLEU |
| Adj | 0.319 | 3.36 | 0.801 | 2.76 | 0.811 | 2.81 |
| Adv | 0.134 | **6.56** | 0.798 | **5.45** | 0.815 | **5.93** |
| Verb | **0.383** | 3.20 | **0.839** | 2.50 | 0.853 | 3.83 |
| Noun | 0.314 | 2.97 | 0.806 | 2.53 | **0.860** | 2.99 |

Table 2. Performance across POS tags for *en*.

| Word length | sgns | | char | | electra | |
|---|---|---|---|---|---|---|
| | cosine | lemma-BLEU | cosine | lemma-BLEU | cosine | lemma-BLEU |
| short | **0.332** | 3.19 | **0.845** | 2.58 | 0.817 | 3.10 |
| medium | 0.314 | 3.19 | 0.842 | 2.74 | **0.867** | **3.41** |
| long | 0.327 | **3.66** | 0.694 | **3.00** | 0.854 | 3.33 |

Table 3. Performances across word lengths for *en*.

## Embeddings and Words




(a) *sgns* no obvious pattern (b) *char* clustered by word length (c) *electra* clustered by word frequency

Figure 3. Visualization of gold embeddings with word labels for English.

### References


Mickus, T., Paperno, D., Constant, M., and van Deemter, K. (2022). SemEval-2022 Task 1: CODWOE – comparing dictionaries and word embeddings. SemEval.
Chen, P. and Zhao, Z. (2022). A unified model for reverse dictionary and definition modelling. arXiv.


### Acknowledgement


Pinzhen Chen is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825303 (Bergamot). Zheng Zhao is supported by the UKRI Centre for Doctoral Training in Natural Language Processing (UKRI grant EP/S022481/1). This work reflects the view of the authors, and not necessarily that of the funders.


THE UNIVERSITY of EDINBURGH informatics

Horizon 2020 European Union Funding for Research & Innovation

bergamot

NLP UKRI CENTRE FOR DOCTORAL TRAINING