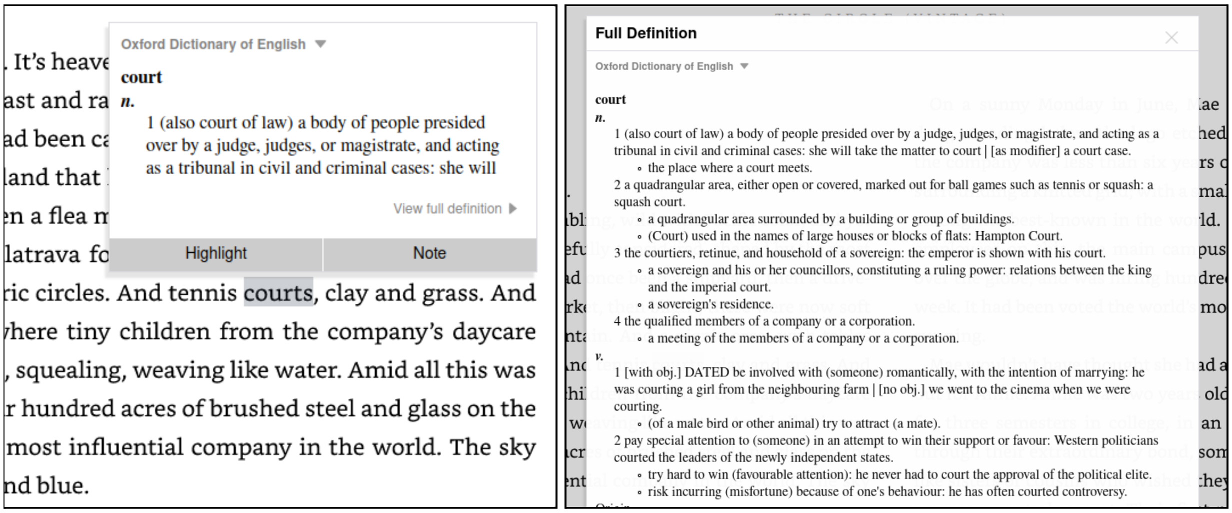
Matching contextual and definitional embeddings for a sense-aware reading assistant

Internship proposal - Carlos Ramisch and Alexis Nasr, SELEXINI ANR project

Imagine you are reading a book in a foreign language that you understand quite well, but you are not totally fluent in. At some point, you come across a word that you do not understand in a sentence. Imagine you can click on the word in your screen and its **definition** shows up (like in the ebook reader shown in Figure 1). A definition is a text snippet that explains the meaning of that word using other words that you are more likely to be familiar with. It allows you to access the meaning of the unknown word and fully grasp the meaning of the whole sentence.

Suppose that the unknown word (e.g. *cell*) has multiple senses (e.g. "*a small room in which a prisoner is locked up"* or *"the smallest structural and functional unit of an organism"*, among others). Instead of a single definition, you will get a list of definitions, and you will have to read through all of them to decide which one is more appropriate in this **context**. As a human, you will probably be able to interpret the known words in the context of the unknown word, and choose the definition that better matches this context.



. It's heave
ast and ra
ad been ca
land that
en a flea m
latrava fo
ric circles. And tennis courts, clay and grass. And
vhere tiny children from the company's daycare
, squealing, weaving like water. Amid all this was
r hundred acres of brushed steel and glass on the
most influential company in the world. The sky
nd blue.

Oxford Dictionary of English ▾

court
n.
1 (also court of law) a body of people presided over by a judge, judges, or magistrate, and acting as a tribunal in civil and criminal cases: she will

Full Definition

Oxford Dictionary of English ▾

court
n.
1 (also court of law) a body of people presided over by a judge, judges, or magistrate, and acting as a tribunal in civil and criminal cases: she will take the matter to court | [as modifier] a court case.
- the place where a court meets.

2 a quadrangular area, either open or covered, marked out for ball games such as tennis or squash: a squash court.
- a quadrangular area surrounded by a building or group of buildings.
- (Court) used in the names of large houses or blocks of flats: Hampton Court.

3 the courtiers, retinue, and household of a sovereign: the emperor is shown with his court.
- a sovereign and his or her councillors, constituting a ruling power: relations between the king and the imperial court.
- a sovereign's residence.

4 the qualified members of a company or a corporation.
- a meeting of the members of a company or a corporation.

v.
1 [with obj.] DATED be involved with (someone) romantically, with the intention of marrying: he was courting a girl from the neighbouring farm | [no obj.] we went to the cinema when we were courting.
- (of a male bird or other animal) try to attract (a mate).

2 pay special attention to (someone) in an attempt to win their support or favour: Western politicians courted the leaders of the newly independent states.
- try hard to win (favourable attention): he never had to court the approval of the political elite.
- risk incurring (misfortune) because of one's behaviour: he has often courted controversy.

Figure 1: ebook excerpt from Amazon's web Kindle reader (https://read.amazon.com/). When clicking on a word (*court*), a definition pops up. Only one (incorrect) definition is shown here. It is also possible to see all 6 definitions from the dictionary. The presence of the context word *tennis* could help disambiguate and show only relevant definitions.

The goal of this internship is to **develop and evaluate an original NLP model capable of aligning a word's context with its correct definition, even if the word is ambiguous**, i.e., having more than one definition listed in the dictionary. To achieve this ambitious goal, the recruited intern will address 3 challenges that constitute major milestones of the project.

1. **Representations:** current NLP models represent words and sentences as real-numbered vectors. A first baseline would be to apply the Lesk algorithm, which counts the number of overlapping words between the context and the definitions (Basile et al. 2014).

Another simple method consists in embedding both the context of the word and its candidate definitions using a pre-trained language model such as BERT (Devlin et al. 2018), returning the definition whose embedding is closest to the context embedding. As definitions are often composed of a genus-differentia pair, their embeddings could be fine-tuned to represent this structure, using techniques for automatic hypernym extraction (Camacho-Collados et al. 2018), hyperbolic spaces (Nickel & Kiela 2017) or graph embeddings that encode the relations between word senses (Nguyen 2020). Semeval 2022 - task 1 CODWOE can also provide useful insights into definition embeddings.

2. **Alignment:** When definitions and contexts are embedded into a shared space, comparing them is trivial. However, if this is not the case, it might be necessary to learn a transformation between these spaces, analogous to cross-lingual embeddings (Lample et al. 2018). Moreover, contextual embeddings such as BERT often represent wordpiece tokens instead of full words (Schuster & Nakajima 2012). Finally, the occurrence of multiword phrases both in texts and dictionaries complicates the alignment further, requiring some technique to identify them in advance (Ramisch et al. 2020).
3. **Evaluation:** traditionally, the task that consists in assigning a sense to a word occurrence is word sense disambiguation (Navigli 2009). Beyond this straightforward possibility, it is possible to evaluate the model using a word-in-context task (Pilehvar & Camacho-Collados 2019) sense-aware similarity datasets (Huang et al. 2012), etc. Since the project's target language is French, evaluation datasets must cover this language.

This internship will take place in the context of the recently funded ANR SELEXINI project. The project aims at developing lexicon induction methods to create a large structured semantic lexicon for French. One of the by-products of this internship is a large French corpus with corresponding contextual embeddings aligned to Wiktionary entries. The intern will join the TALEP team in Luminy, Marseille, and have the opportunity to interact with researchers in the partner universities (Univ. de Saclay, Univ. de Paris, Univ. de Lorraine) and submit a paper to an international conference, depending on the results of the internship.

The development of a showcase web interface for clicking and retrieving definitions is a possible extension after the internship (e.g. as a summer project). Other possible extensions include generating definitions on the fly for words or senses not present in Wiktionary (Bevilacqua et al. 2020), or generating definitions for whole phrases and multiword expressions. These can be the topic of a PhD thesis if the intern shows interest and skills compatible with the project.

Requirements:

- Familiarity with word embeddings
- Python programming is highly recommended
- Fluent English reading, reasonable English writing
- Curiosity, autonomy, rigour, scientific methodology
- Interest in linguistics, languages, dictionaries, semantics