CHAPTER 1

Introduction

1.1 RATIONALE FOR WORKING WITH COMPARABLE CORPORA

The ability of the computers to handle larger amounts of texts and the availability of more texts in electronic form led to the rise of data driven research in computational linguistics. In the case of Machine Translation (MT) and other kinds of multilingual Natural Language Processing (NLP), the first source of large data came from collections of translations, initially in the Statistical MT (SMT) approach from IBM [Brown et al., 1990], which was based on the Proceedings of the Canadian Parliament in English and French as their data source. This research direction was followed by a proliferation of SMT models, which relied on larger and larger collections of *parallel* data, which consist of exact translations between a pair of languages or several languages at the same time.

However, this early research also demonstrated the limits of using fully parallel corpora, with the most pressing initial concern on the amount of such data [Fung, 1998, Rapp, 1995]. This led to another strand of studies concerning the use of *less* parallel sources of texts, usually under the name of comparable corpora [Sharoff et al., 2013]. This book will present an overview of the modern approaches to building and using comparable data in multilingual NLP.

We will start by outlining the basic principles of using comparable resources as well as by comparing them to fully parallel resources (this chapter). This will be followed by specific chapters on building comparable corpora (Chapter 3), aligning their sentences to find a database of suitable translations (Chapter 4), using these corpora to produce dictionaries and termbanks (Chapter 5), to build MT engines (Chapter 6) and to use them in other applications (Chapter 7).

1.1.1 AVAILABILITY OF TRULY PARALLEL DATA

The first limitation in the use of parallel corpora comes from the process of their production. For truly parallel corpora we need to collect texts that have been carefully translated by highly trained professional translators [Massey, 2017]. Many more people produce monolingual texts in their native languages in comparison to output of a small number of trained translators. Also there is an imbalance in the amount of translations produced for a relatively small number of major languages, primarily for the languages of the United Nations or the EU, while there are thousands of languages with very minor resources. Less resourced languages can still benefit from parallel resources. However, statistically speaking, language can be described as a large number of rare events in the sense that an individual word or expression is relatively infrequent, but the totality of rare events contributes to the mass probability of words and expressions in a text [Baayen, 2008]. This creates

problems of sparsity even for better-resourced languages. For example, the word *unicyclist* occurs 84 times in 2 billion words of a monolingual English ukWac corpus [Baroni et al., 2009], i.e., about once per 23 million words in ukWac, with no examples of this word in the English parts of large publicly available parallel corpora such as Europarl [Koehn, 2005] or the United Nations corpus [Ziemski et al., 2016].

Another constraint on corpus data concerns the availability of translation products, as it is easier to obtain translated texts produced by large public bodies, such as the European Parliament or the United Nations than many other kinds of translations. This bias leads to many word choices, which are specific with respect to the genres and topics available in such corpora. For example, there are 75 occurrences of the expression *strong voice* in Europarl, all of which are used in the sense of political authority, for example, *ensuring that smaller Member States retain a* ***strong voice*** *in the decision-making procedures*. At the same time, out of the 28 occurrences of this expression in the British National Corpus [Aston and Burnard, 1998] 19 examples refer to the quality of human voices, for example, *She had a good,* ***strong voice*** *– an actor's voice*. When translating *strong voice* into other languages, the political authority metaphor needs to be explicitly unpacked, because the literal translation is not suitable. For example, when translating *strong voice* into Russian, the political sense is likely to use *reshitelno vystupatj* ('to express assertively') or *goryacho osuzhdat* ('to condemn vehemently') as opposed to the straightforward literal translation *gromkij golos* ('loud voice'). In the end, the mismatch between the domains and genres of parallel corpora and the target applications can lead to errors, which might be corrected by the use of less parallel resources.

1.1.2 TRANSLATIONESE IN PARALLEL DATA

Another kind of problems concerning the use of parallel data comes from a particular phenomenon known as translationese, namely a difference between features of translated texts and texts originally produced by native speakers [Koppel and Ordan, 2011]. Translationese is caused by factors inherent in the translation process, such as explicitation [Frankenberg-Garcia, 2009], i.e., the need to provide more information in the translated text for what remains implicit in the source text. For example, translations tend to use more cohesive markers, such as *therefore, however, nevertheless*, in comparison to original texts by making the logical relations more explicit in translation [Koppel and Ordan, 2011]. Other factors leading to translationese are related to the temporal and cognitive pressures on the translation operations, as the translators need to complete their tasks in a short period of time. This leads such phenomena as (1) normalisation, i.e., the tendendency to re-use stock expressions of the target language even when the source text deviates from the norm, and (2) "shining-through" [Rubino et al., 2016], i.e., the influence of the syntactic and lexical choices stemming from the source texts even when other choices are common in monolingually produced texts in the target languages. In the end it has been shown that translationese effects have statistical significance leading to the ability to build fairly accurate classifiers detecting texts translated by humans [Rabinovich and Wintner, 2015], even in an unsupervised fashion [Riley et al., 2020].

From the viewpoint of the translation direction, there is always a source text which needs to be translated into other languages, so only one text in a parallel corpus is primary, while other texts exhibits features of translationese. At the same time, when parallel corpora are used for MT applications, this directionality of translation is usually ignored, so that the Slovenian→ English MT built from the Europarl corpus uses slightly unnatural Slovenian texts exhitibing features of translationese as its source texts. These aspects call for the greater use of texts originally produced in the respective languages.

Table 1.1: Adena culture example

en	The Adena culture was a Pre-Columbian Native American culture that existed from 800 BC to 1 AD, in a time known as the Early Woodland period.
de	Adena-Kultur ist die Bezeichnung für eine im mittleren Ohio-Tal ansässige prähistorische Indianerkultur. Sie lässt sich für die Zeit von etwa 1000 v. Chr. bis 200 n. Chr. nachweisen.
fr	La Civilisation Adena était une culture pré-colombienne amérindienne ayant existé de l'an 1~000 à l'an 200 avant J.-C., durant l'ère connue sous le nom de Période sylvicole.
ja	アデナ文化（Adena）は、アメリカ合国オハイオ州を中心に1000B.C. から元前後にえた文化。アデナ文化は、初期（ないし前期）ウッドランド期（Early Woodlannd Period）の文化として位置付けられ、アデナ文化の出によって、後のホプウェル文化をはじめとするウッドランド文化の先となるウッドランド式土器や丘墓、トウモロコシ耕のすべてがでそろった。
ru	Культура Адена — доколумбовая индейская археологическая культура, существовавшая в период 1000—200 г. до н. э., в период, известный как ранний Вудлендский период.

1.2 LEVELS OF COMPARABILITY

These constraints on the availability and the use of parallel data led to numerous studies utilising less parallel resources. This research direction generally goes under the name of 'comparable corpora'. However, it is important to note that there is no clear dividing line between fully parallel and comparable corpora, as multilingual resources vary with respect to the degree of linking between documents in the two languages. Consider some examples of documents along the cline of comparability:

translations identifiable originally produced source texts and their translations.

truly parallel true high quality translations, such as the proceedings of the European Parliament.

modified parallel translations with some modifications to cater for the target audience. For example, language-specific descriptions of the Search dialogue box in the OpenOffice manual are translations from English with necessary modifications, such as searching for *New York* is replaced with searching for *Berlin* in the German version.

adaptations translators exhibit freedom in rendering the source text, as it is the case with many of the fan-produced subtitles as in the OpenSubtitles corpus [Tiedemann, 2012]

strongly comparable closely related texts produced in several languages.

Wikipedia entries entries on exactly the same topic are linked across languages via iWiki links, see Table 1.1, with individual entries varying in the amount of information.

news items very specific events are covered by numerous news agencies in various languages. Often the same agency reports the same story in various languages without fully relying on their translation, see the BBC News in English and Spanish.

weakly comparable similar texts which cannot be directly linked to each other across languages, while still being in the same domain and genre, for example:

- texts in the same narrow subject domain and genre, but describing different events, e.g. parliament debates on health care from the German Bundestag, the British House of Commons and the French parliament;
- texts within the same broader domain and genre, but varying in subdomains and specific genres, e.g. parallel queries in the renewable energy domain mostly returning wind energy research articles in English vs solar panel producers in Russian [Sharoff, 2013].

unrelated collections of unrelated texts, which are nevertheless collected using comparable methods from comparable sources. For instance, this concerns the use of random snapshots of the Web for Chinese, English, German and Russian [Sharoff, 2006] or the use of filtered Common Crawl data [Wenzek et al., 2019] with the assumption that different cultures use the Web for broadly similar purposes.

Further on in the book we will refer to collections of documents on this comparability scale as parallel, strongly comparable, weakly comparable and unrelated corpora.

1.3 METHODOLOGY FOR DEALING WITH COMPARABLE RESOURCES

Even though there is no clear dividing line between parallel and comparable resources, there are two clearly distinct ways of using such resources, which focus on the amount of translation information present in each resource.

The first approach assumes the possibility of obtaining very local information which can link two languages through splitting corpora into small translation units, which are supposed to convey the same meaning in both languages. Usually this is done on the sentence level. This is the classic case of training Machine Translation (MT) engines, as MT naturally assumes a high level of parallelism as the starting point. This can be also considered as a fully supervised model of translation.

The second approach assumes building global multi-dimensional spaces for the entire set of linguistic phenomena in each language to determine the similarity between them. This is the classic case of using comparable resources for dictionary induction, the phenomena in this case are words or other kinds of dictionary entries. As usual in computational linguistics there are many studies blending the local and global approaches.

Irrespectively of choosing the local or global approach to using corpora, we need to determine ways of measuring similarity. Modern approaches usually rely on the Vector Space Model (VSM), which uses a numerical vector x to represent any entity, such as a word, a phrase, a sentence or a link between entities. If a VSM assign similar vectors to entities, which are translations of each other in the respective languages, this shared space can be the basis for identifying translations via comparable corpora. We will start with an introduction into the basic principles of building cross-lingual VSMs, see Chapter 2, and then we will proceed to methods for utilising these spaces, first for finding similar documents, see Chapter 3, then for finding translations on the sentence level, see Chapter 4 and finally for bilingual dictionary induction on the word level, Chapter 5. The shared VSM space is also a natural starting point for developing Machine Translation engines with no or with limited parallel data, see Chapter 6.

The utility of comparable corpora goes beyond applications in translation. As the amount of resources required for training NLP tools for text annotation, document classification or retrieval is limited in many languages, it is possible to achieve low-resource NLP via few-shot or zero-shot models. This means an NLP model can be trained for a low-resource language with very few or even no annotated examples in this language provided that we have a sufficient number of annotated examples in another (donor) language and that we have comparable corpora for both languages. Methods for making cross-lingual predictions in this context are discussed in Chapter 7.

CHAPTER 2

Basic principles of cross-lingual models

When we start working across languages, we need to determine ways of measuring similarity of entities (such as a word, a phrase, a sentence or a link between entities) within each language and across languages. Modern approaches usually rely on the Vector Space Model (VSM), which uses a numerical vector x of a specified dimensionality D to represent any entity. The entities are similar when their vectors are not too far apart.

2.1 MONOLINGUAL VSMS

Early VSM approaches [Salton et al., 1975] constructed these vectors in an intuitive model by computing the strength of association between entities. For example, a word can be represented as a vector of other words it co-occurs with. The values of these vector can be either normalised frequencies of co-occurences or some kinds of collocation scores in a window of a certain size [Manning and Schütze, 1999]. Table 2.1 lists example vectors for three words *moon, Sun* and *zebra*.[1]

Table 2.1: Examples of word vectors in an association matrix

	crescent	cross	full	giraffe	half	hot	ibm	moon	rise	see	shine	stripe	sun
moon	9.3	0.0	8.2	0.0	7.9	0.0	0.0	0.0	6.8	3.3	7.6	0.0	10.5
sun	0.0	0.0	5.3	0.6	0.8	8.7	8.3	10.5	8.1	4.1	11.1	0.0	0.0
zebra	0.0	13.6	0.0	9.3	0.0	0.0	0.0	0.0	0.0	0.7	0.0	11.8	0.0

Each row of this matrix $\mathbf{M}$ stores the Point-wise Mutual Information scores [Church and Hanks, 1990] for the collocates:

$$PMI = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \tag{2.1}$$

where p is the probability of seeing words w_1, w_2 or their collocation "$w_1 w_2$" in this corpus. The PMI value is assumed to be zero if $p(w_1, w_2)$ is zero, i.e. if the two words do not co-occur in a specific window, the distance of three words in either direction was chosen for computing this table.

This similarity of vectors implies the similarity of meanings according to the hypothesis of distributional similarity, *You shall know a word by the company it keeps*, as originally formulated

[1]The frequencies have been taken from the British National Corpus [Aston and Burnard, 1998], which was compiled in the beginning of the 1990s. This timeframe leads to the specific *Sun-IBM* collocation.

by Firth [1957], see also [Harris, 1954, Turney and Pantel, 2010], i.e., semantically similar words are likely to have similar vectors of co-occurences.

A VSM representation of this kind is interpretable, as one can see how words are related to each other via their collocations, so that we can see the conditions under which words are similar, for example, we can see that *moon* and *sun* share a number of collocates, and under which they differ, because the collocations apply to one of the senses but not to others. For example, some collocations are specific to one sense of *sun* (*hot, moon, rise, shine*), while others are specific to another sense, such as its collocation with *IBM* for the computer company sense of *Sun*, as in the BNC example *offered neither by Sun nor IBM*.

Even though the model is easily interpretable, its application creates a problem with the number of dimensions in the co-occurrence vector, because the vector for each word needs to include *all* words selected as possible collocates for every other word in this corpus. This becomes intractable in a large corpus with a lexicon easily exceeding several million words, while the matrix of collocations will be the square of this size. Depending on how aggressively one reduces the number of the most important keywords (this defines D, the dimensionality of each row in matrix $\mathbf{M}$), the number of keywords in a large corpus is nevertheless likely to exceed 100,000. This is still not practical, as there are many issues associated with high-dimensional spaces, such as the curse of dimensionality and hubness, when some entities are more likely to become closer to a large number of other entities, thus polluting the usefulness of similarity estimates. The likelihood of such problems rises with the number of dimensions [Radovanović et al., 2010]. At the same time, the matrix is very sparse, as most of its values are zeros. For example, *zebra* does not co-occur with most of the collocates of *moon* and *sun*, while it adds its own set of collocates *giraffe*, *stripe*, and *cross* (the last collocate comes from the sense of *zebra crossing*). In the end, the vector for *zebra* is similar to vectors for other animals and their properties (*giraffe* and *stripe*), as well as to various concepts related to the transport infrastructure (from its sense of *zebra crossing/*), which in turn have very few collocations shared with *moon*.

A VSM space can be compressed to a denser representation by keeping only a smaller number of dimensions $d << D$ which still retains most of the variation available in the full matrix of co-occurrences. A simple way to achieve this is by using the Singular Value Decomposition of matrix $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ [Rapp, 2004] to derive a low-rank approximation matrix $\mathbf{M}_d$ such that:

$$\mathbf{M}_d = \mathbf{U}\mathbf{\Sigma}_d\mathbf{V}^T \tag{2.2}$$

in which only the largest d singular values of $\mathbf{\Sigma}$ are retained. The commonly used value of d is between 50 and 500. Another popular approach for building dense VSM spaces is Random Semantic Indexing (RSI), see [Sahlgren et al., 2008], which also avoids the problem of making a direct SVD transform on a co-occurrence matrix with millions of dimensions. Performing either SVD or RSI produces non-sparse vectors of a reasonable number of dimensions, which simplifies many operations in this vector space. However, the downside is that the resulting dimensions no longer interpretable.

The neural approach to building VSMs is similar to the collocation-based approach except that it replaces the need of counting the co-occurrences with predicting their probability. If we take a very large set of examples in which one word has been removed, such as

Examples 2.1 *shine of the .... upon the harbour*

Examples 2.2 *light of the .... shining between clouds*

it is possible to recover the most likely masked word with a certain degree of precision, for example, for (2.1) and (2.2) this is likely to be either *moon* or *sun*. Therefore, we can learn dense vector representations (called embeddings in the neural approach) which assign similar vectors to words which can be used in similar contexts. For example, the initially random vectors for *sun* and *moon* through this training are becoming similar, because their embedding vectors are trained to predict similar contexts, such as *full, rise* or *shine*. From a random initialisation state, a neural network is trained to learn more suitable embedding vectors for the respective words with backpropagation from the masked word prediction task. For the best results we need to adjust the schedule of presenting words (to balance the frequency effects) and to choose the rate of negative examples, i.e. words not occurring in the chosen context by randomly inserting them into the empty slot. Randomly chosen words will either violate grammatical constraints or offer unreasonable contexts:

Examples 2.3 *shine of the APPLY|BLOOD upon the harbour*

More formally, we can build word embedding representations by modeling the probability of seeing a word in the context of m other words:

$$p(w, c_{1:m}) = \frac{1}{1 + e^{-\sum w \cdot c_i}}$$

where the function is the sigmoid, which assumes the probability close to 1 for real corpus examples and the probability close to 0 for negative examples. We can find the best vectors w and c to match the probabilities for the positive and negative examples by using the Stochastic Gradient Descent algorithm, see Chapter 10 in [Goldberg, 2017] for more information on the training process. Learning representations rather than counting co-occurrences usually leads to better generalisation capabilities of the neural embedding models in comparison to the count-based ones [Baroni et al., 2014].

The beauty of neural methods is that embedding representations can be learned on the levels of words, sentences, paragraphs or documents, thus providing vector representations, which are suitable for many downstream tasks, so that embeddings can be used for Machine Translation, text classification, natural language inference or part-of-speech tagging. The objective of learning suitable embeddings aims at producing similar vectors (i.e. close with respect to their distance in the VSM space) for words, sentences, etc, when they are used in similar contexts, therefore similar embedding vectors are assumed to be similar in meanings. The common practice is to use embedding

VSMs with around 300 dimensions for word embedding models (same as with SVD or Random Semantic Indexing), rising to around 1000 dimensions for sentence or paragraph embedding models.

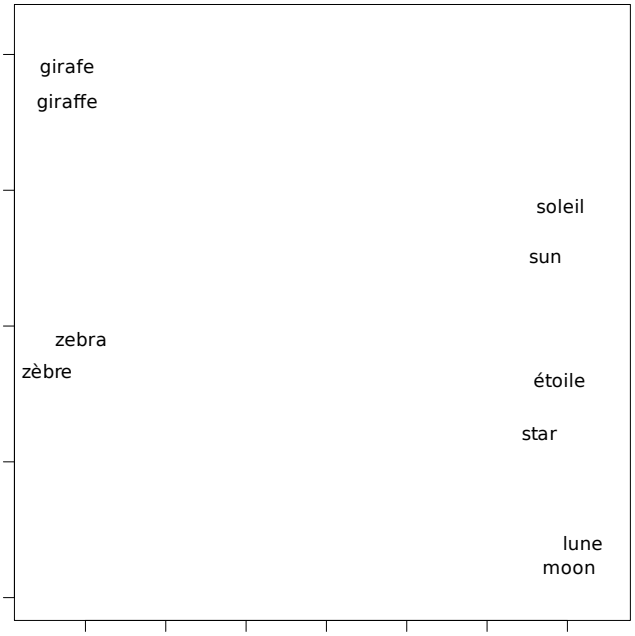

Figure 2.1: Example of aligned word spaces

2.2 CROSS-LINGUAL VSMS

Creating VSMs that can encode multiple languages is very useful in a multilingual context. The VSM spaces can be either built in a way which creates them shared across several languages, or two VSM spaces from unrelated comparable corpora can be aligned post-hoc, for example, by using a seed dictionary or merely identical tokens, such as numbers or punctuation marks. In the case of post-hoc alignment, the two independent VSM representation spaces are transformed with the objective of allocating the seed dictionary elements as close as possible in this aligned space.

We can use comparable corpora in building VSM representationss if we can find a way of aligning the VSMs for several languages, which means obtaining similar vector representations for similar entities (words, constructions or sentences) **across** languages. Figure 2.1 shows word vectors for several English words (*moon, sun, giraffe, zebra*) and the related French words (*lune, soleil, girafe, zèbre*) according to the aligned word embeddings of fasttext [Bojanowski et al., 2017], as obtained after reducing the dimensionality of the $d = 300$ fasttext vectors to a two-dimensional plot

via a t-SNE transformation [Van der Maaten and Hinton, 2008]. We can determine that the closest French word to *sun* in this space is *soleil*, which helps in inferring their semantic similarity.

2.3 CONTEXTUAL EMBEDDINGS

One of the problems with traditional vector space models is the need to represent the meaning of each word as a single vector. Such examples as *sun* and *zebra* show that the assumption that words with similar vectors have similar meanings is often difficult to satisfy. Many words have very different meanings in different contexts, so that the embedding for *sun* needs to be close to both *moon* and *IBM* at the same time, while this is not true for its likely French translation *soleil*.

A way of representing the way how the embeddings for words are combined together to produce meanings in their contexts is offered by means of connected layers of mathematical objects called artificial neurons (they are called neurons because their design was inspired by biological neurons, even though the objects themselves and their networks are usually quite different from the biological neurons). Three relevant architectures here are (1) Recurrent Neural Networks (RNNs), which model how much the embedding for the next word adds to the embedding vector obtained from the preceding words, (2) Convolutional Neural Networks (CNNs), which model patterns of co-activation of word embeddings irrespectively of their order, and (3) Transformers, which model the so called attention mechanism, i.e. how words influence each other. For example, for a sentence like

Examples 2.4 *The **food** was **amazing**, **but** the **service left** something to be **desired**.*

the transformer builds attention links between words, such as the objects and their properties, e.g., *food* and its quality. The weights for combining word embeddings in all of these architectures are estimated by means of back-propagation of errors in a classification task, i.e., the weights are optimised to reduce the classification loss errors using such algorithms as Stochastic Gradient Descent. The principles behind neural architectures and their training are introduced in far greater details in [Goldberg, 2017].

Contextual embedding models, such as ELMO or BERT, combine the mechanism of training word embeddings to predict masked words, such as in Examples (2.1) or (2.2) with the mechanism of neural networks which model how words can influence each other. For example, in the case of BERT, the transformer model is pre-trained on a large corpus (such as Wikipedia) to predict masked words or whether a sequence of two sentences is real or not. This pre-trains the weights for attention links from a very large raw text corpus without any explicit annotation. The attention links which can be detected as the outcome of this translating can already predict many kinds of syntactic and semantic relations, for example, subject-verb agreement, even without any explicit training [Rogers et al., 2020]. The weights can be later "fine-tuned" to a downstream task by training on a downstream corpus, for example, the sentiment of the whole sentence or evaluation of its specific properties, such as food or service [Pontiki et al., 2016].

The weights in each layer of the pre-trained transformer models determine how the embeddings for individual words are modified dynamically for each example. In the end, the embedding for *Sun* in the upper layer of this architecture when used in a context like *offered neither by Sun nor IBM* will be different from the embedding the same token in the same layer when used in a context like *alignment of the Sun, Moon and Earth*. The pre-trained transformer models started with BERT [Devlin et al., 2019] and this was later followed by a family of models, such as Roberta [Liu et al., 2019] or Distilbert [Sanh et al., 2019].

Even in the absence of truly parallel corpora or any dictionary seeds (as used for producing the representations like Figure 2.1, pre-trained language models can achieve a good degree of parallelism through unsupervised methods, as they detect inherent parallelism in the distribution of frequencies of words and constructions via sharing the weight parameters [Conneau et al., 2020, Pires et al., 2019]. For example, if we take Wikipedia corpora for English and French and set the task of predicting the masked words for a number of sentences like

Examples 2.5 *Consult the TABLE beam sizes below* vs
Vous pouvez consulter le TABLEAU des rémunérations des professeurs
'you can consult the TABLE of the salaries of the teachers'

where the capitalised *TABLE* and *TABLEAU* are the masked words, the resulting embedding representations for many words in the lexicon will be sufficiently close to their relevant translations in another language, even though the sentences themselves are not parallel translations.

In the remaining chapters of this boook, we will show how representations of this kind are useful to predict the similarity of documents, sentences and words across languages, as well as to build MT and other applications.

We want to stress here that the majority of methods for building multilingual VSMs rely on comparable corpora with the most successful recent methods relying on unrelated comparable corpora. Nevertheless, to understand which models are more successful than others we need to understand how comparable their pre-training corpora are. For example, the multilingual BERT is trained on Wikipedias, while XLM-Roberta is trained on the multilingual Common Crawl corpus [Wenzek et al., 2019]. Sharing parameters in their pre-training comes from the comparability of their underlying corpora, which we are going to investigate in the forthcoming chapters.