Natural Language Engineering (2024), **30**, pp. 150–170
doi:10.1017/S1351324922000134

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Real-world sentence boundary detection using multitask learning: A case study on French

KyungTae Lim[1,†] and Jungyeul Park[2,3,*,†]

[1]Hanbat National University, Daejeon 34158, South Korea, [2]The University of British Columbia, Vancouver, BC V6T 1Z4, BC, Canada, and [3]University of Washington, Seattle, WA 98195, USA
*Corresponding author. E-mail: jungyeul@mail.ubc.ca

(Received 29 March 2021; revised 6 March 2022; accepted 7 March 2022; first published online 6 April 2022)

Abstract

We propose a novel approach for sentence boundary detection in text datasets in which boundaries are not evident (e.g., sentence fragments). Although detecting sentence boundaries without punctuation marks has rarely been explored in written text, current real-world textual data suffer from widespread lack of proper start/stop signaling. Herein, we annotate a dataset with linguistic information, such as parts of speech and named entity labels, to boost the sentence boundary detection task. Via experiments, we obtained F1 scores up to 98.07% using the proposed multitask neural model, including a score of 89.41% for sentences completely lacking punctuation marks. We also present an ablation study and provide a detailed analysis to demonstrate the effectiveness of the proposed multitask learning method.

Keywords: Sentence boundary detection; French; Multitask learning; Corpus creation

1. Introduction

Sentence boundary detection (SBD) is a basic natural language processing (NLP) task that detects the beginning and end of the sentence. Previous works (Palmer and Hearst 1997; Reynar and Ratnaparkhi 1997; Kiss and Strunk 2006; Gillick 2009; Lu and Ng 2010) considered sentence boundary disambiguation as a classification problem. Past researchers classified full-stop punctuation marks and abbreviations to determine the ends of sentences. Note that sentence boundary disambiguation differs from SBD because the former requires punctuation marks for classification, whereas the latter does not necessarily require them to determine the boundary of the sentence. Hence, we use the term "disambiguation" without the acronym. We use SBD explicitly for the sentence boundary detection task. To illustrate the problem with previous approaches for sentence boundary disambiguation, we evaluated a simple paragraph (see Figure 1) using state-of-the-art systems (e.g., SSPLIT in CoreNLP (Manning *et al*. 2014), ELEPHANT (Evang *et al*. 2013), and SPLITTA (Gillick 2009)) for English.

The paragraph in Figure 1 contains five sentences and fragments, for which humans can easily exploit their linguistic competence to detect the sentences. In the paragraph, there are two noun fragments (*Opening of the session* and *Agenda*) and three complete sentences containing a subject and a predicate. Although the complete sentence ends with a period, the noun fragments do not contain proper punctuation. All three state-of-the-art systems failed to identify the correct sentence boundaries, detecting only three of the five sentences based on the punctuation marks. Without punctuation marks representing the ends of sentences, the systems identified two noun

[†]KyungTae Lim and Jungyeul Park contributed equally.

```
Opening of the session

I declare resumed the 2000-2001 session of the European Parliament.

Agenda

Mr President, the second item on this morning's agenda is the recommendation for
second reading on cocoa and chocolate products, for which I am the rapporteur. Quite
by accident I learnt yesterday, at 8.30 p.m., that the vote was to take place at
noon today.
```

Figure 1. Paragraph example from the Europarl corpus (Koehn 2005).

fragments as parts of the following sentences. Therefore, another method to achieve SBD relying on features at the beginnings of sentences is required. This is a challenging problem because it must handle the non-appearance of punctuation marks in addition to capitalized words, such as *I* or *Mr*. These terms use capital letters even in the middle of a sentence. The French language shows similar characteristics as English for SBD. Some words starting with a capital letter, such as *Monsieur* ("Mr"), can also appear in the middle of a sentence.

In this study, we apply SBD to written French text: an approach that has rarely been explored in the literature. The objective of this paper is to leverage linguistic information, such as parts of speech (POS), named entity recognition (NER), and capitalized words to identify the beginning of a sentence instead of classifying the end of the sentence using punctuation marks. Our main contributions are three-fold. First, an effective method to construct SBD on a modern corpus is provided to solve common sentence-marking deficiencies. Second, a multitask learning approach is presented that predicts linguistic information (e.g., POS and NER) by training sentence boundaries simultaneously, as in a real-world setting. Third, the effects of multitask learning are explored wherein the number of training data and the multitask procedures vary.

We first present previous works in Section 2; then, we construct training and evaluation datasets for French in Section 3. We then propose our novel approaches for SBD using sequence labeling algorithms with multitask learning, including baseline conditional random field (CRF) and contextualized neural-network (NN) models discussed in Section 4. Thereafter, we report on our SBD experiments and their results, including a comprehensive discussion of our model's application in a real-world setting. These are discussed in Sections 5 and 6. We finally present a conclusion in Section 7.

2. Previous works

Most previous works (Palmer and Hearst 1997; Reynar and Ratnaparkhi 1997; Kiss and Strunk 2006; Gillick 2009) considered sentence boundary disambiguation as a classification problem in which they classified full-stop punctuation marks and abbreviations ending with a period to find the end of a sentence. More recently, Evang *et al.* (2013) developed a character-level classification system for tokenizing words and sentence boundaries. Although most previous works sought to identify the ends of sentences, Evang *et al.* (2013) and Björkelund *et al.* (2016) detected their beginnings. Table 1 summarizes some previous works' approaches to sentence boundary disambiguation.

Sentence boundary disambiguation and SBD have rarely been explored for languages other than English. We address the French language in this paper. González-Gallardo and Torres-Moreno (2017) tackled a sentence boundary disambiguation problem wherein a binary classification task was implied. Azzi, Bouamor, and Ferradans (2019) applied sentence boundary disambiguation to text from scanned portable data files in both English and French, excluding the

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

Table 1. Summary of previous works on sentence boundary disambiguation.

PH1997	Classified punctuation marks based on preceding and following POS labels for English and transferred the system for French and German
RR1997	Used the maximum-entropy Markov model to classify punctuation marks for English
KS2006	Introduced an unsupervised language-independent sentence boundary disambiguation system using collocation detection to build an abbreviation detector
G2009	Used a support vector machine with Naive Bayes for periods to distinguish between sentence boundary and an abbreviation in English
LN2010	Performed punctuation prediction on speech utterances using dynamic conditional random fields (CRFs) for Chinese and English
EAL2013	Used CRFs at the character level for word and sentence segmentation with a character-embedding vector for Dutch, English, and Italian
BAL2016	Trained a dependency parser for joint sentence boundary disambiguation and parsing for English

PH1997 (Palmer and Hearst 1997), RR1997 (Reynar and Ratnaparkhi 1997), KS2006 (Kiss and Strunk 2006), G2009 (Gillick 2009), LN2010 (Lu and Ng 2010), EAL2013 (Evang *et al.* 2013), and BAL2016 (Björkelund et al. 2016).

```
Les Sables-d'Olonne La Chaume Philippe et Véronique, son neveu et sa nièce ; Anne,
Albert et Chantal, son beau-frère et sa belle-soeur, ont la tristesse de vous faire
part du décès de Monsieur Serge survenu le 26 novembre 2014, à l'âge de 82 ans. ...
```

Figure 2. Raw SBD data for French: (translation) "*Les Sables-d'Olonne La Chaume Philippe and Véronique, his nephew and niece; Anne, Albert and Chantal, his brother-in-law and sister-in-law, are sad to report the death of Mr. Serge, who passed away on November 26, 2014, at the age of 82. . ..*"

noisy parts. Apart from these previous works, Read *et al.* (2012) described nine available systems and their benchmarks to define standard datasets for evaluation. Dridan and Oepen (2013) discussed document parsing by focusing on evaluation methods for tokenization and sentence segmentation from raw string inputs. In the domain of automatic speech recognition, several SBD-related works have been proposed (Treviso, Shulby, and Aluísio 2017; González-Gallardo and Torres-Moreno 2018). However, because we propose novel approaches for detecting sentence beginnings without relying on punctuation marks in written text, we focus on the detection aspect: SBD.

For a neural system, Xu *et al.* (2014) implemented a hybrid NN-CRF architecture to detect sentence boundaries from audio transcripts. Qi *et al.* (2018) considered joint tokenization and sentence segmentation as a unit-level sequence tagging process based on an NN model. SBD systems that used this approach have achieved the best performance over the last few years. More recently, bidirectional encoder representations for transformers (BERT)-like models (Liu et al. 2019; Martin et al. 2020; Conneau et al. 2020), which are deep contextualized vector representation methods for a token, have shown outstanding performance in several NLP tasks. Because BERT (Devlin *et al.* 2018) is a language model (LM) that learns from a large quantity of raw texts, such as Wikipedia and news sites, it has the ability to capture contextual information for handling unknown words (Lim *et al.* 2020).

3. Creating an SBD corpus

For the raw text dataset, we crawled obituaries from more than 10 newspapers, including *Le Figaro* and *Ouest-France*. An example of sentences is provided in Figure 2. Note that these sentences usually contain significant numbers of proper nouns, such as person and location names. They

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

can serve as a corpus for information extraction and entity linking in future works. We collected 8725 files containing more than 1-M tokens.

3.1 Tokenization

First, we used the preprocessing tools from MOSES (Koehn *et al.* 2007) for normalizing punctuation marks and tokenization. However, there are several issues and errors in French tokenization. We corrected the tokenization results, including identifying pre-defined entities, such as telephone numbers (*02␣31␣77␣01␣16* → *02-31-77-01-16*), times (*12h␣45* → *12h45*), and web addresses. A binary operator ␣ represents a whitespace delimiter in the written text. We manually identified these patterns and constructed regular expressions for post-processing.

3.2 POS tagging and rough sentence boundaries

Second, POS labels and "rough" sentence boundaries were handled by TREETAGGER (Schmid 1994) based on added punctuation marks, which is required for NER in the next step. Because TREETAGGER only classifies punctuation marks for the ends of sentences, we refer to TREETAGGER'S sentence boundaries as "rough" sentence boundaries.

3.3 NER

Third, because the corpus contains a significant number of proper nouns, named entity labels were assigned. We used NER data provided by Europeana newspapers for French (Neudecker 2016) and trained them with NEURONER (Dernoncourt, Lee, and Szolovits 2017), which implements a bidirectional long short-term memory recurrent NN. We evaluated NER models for French using various sequence labeling algorithms and improved the NER results using semi-supervised learning (Park 2018). To refine the current NER task, we added geographical entities using a list of communes in France,[a] which improved the NER results. The NER-processed corpus on the left side of Figure 3 shows geographical entities (third column, using beginning–inside–outside (`bio`) annotation using a geographical entity dictionary) and RNN-annotated entities (fourth column, with using the inside–outside (`io`) format instead of using the `bio`, owing to the original annotation of the NER corpus, which introduces the `io` format). In the corpus, we removed the surnames to maintain anonymity. We compared two entities via geographical entity assignment and RNN labeling and selected the more pertinent entity. When assigned entities were different, the geographical entity was selected if its length (the number of tokens for the entity) was > 1; otherwise, the RNN-annotated entity was assigned. For example, *Chaume* (a village attached to the city of *Sables-d'Olonne*) was annotated as I-LOC by geographical entity assignment and I-PER by RNN labeling. We selected "I-LOC" because the length of the geographical entity assignment was greater than one. We removed B annotations of the geographical entity assignments to retain only `io` annotations of named entities for the consistency of NER labels (see the right side of Figure 3).

3.4 Marking SENT labels and a manual correction

Finally, we created heuristic rules to determine whether a fragment (e.g., a noun phrase) can be an independent sentence (e.g., geographical entities at the beginning of the text). However, heuristic rules to mark SENT labels (for the beginning of a sentence) were weak, and we manually verified all sentences to correctly mark SENT labels. During manual verification, we corrected SENT labels in addition to incorrectly assigned POS and NER labels, which were initially automatically

[a] https://www.data.gouv.fr/fr/datasets/base-officielle-des-codes-postaux.

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

NER processed:	SENT marked:	
Les DET:ART B-LOC O	Les DET:ART I-LOC B-SENT	'The'
Sables-d'Olonne NAM I-LOC I-LOC	Sables-d'Olonne NAM I-LOC O	'Sables-d'Olonne'
La DET:ART B-LOC O	La DET:ART I-LOC O	'the'
Chaume NAM I-LOC I-PER	Chaume NAM I-LOC O	'Chaume'
Philippe NAM O I-PER	Philippe NAM I-PER B-SENT	'Philippe'
et KON O O	et KON O O	'and'
Véronique NAM O I-PER	Véronique NAM I-PER O	'Véronique'
, PUN O O	, PUN O O	','
son DET:POS O O	son DET:POS O O	'his'
...	...	'...'

Figure 3. Preprocessed SBD data for training (SENT marked): *Les Sables-d'Olonne La Chaume* and *Philippe et Véronique, son* . . . are annotated as two sentences where the latter represents the sentence *middle* that no punctuation marks precede.

assigned using heuristic rules. Errors in POS labels are often caused by words that start with a capital letter in the middle of a sentence. For example, *Très* ('very') is automatically labeled as a proper noun in *Remungol*, *Guénin Plounévézel Très touchée par* . . . ('LOC, LOC LOC very touched by ... '). If we can detect the correct beginning of the sentence in which a fragment, *Remungol* ('LOC'), *Guénin Plounévézel* ('LOC LOC'), is considered to be an independent sentence, and the following sentence starts from *Très* ('very'), the word *Très* would be correctly labeled as an adverb. There were person- and location-label errors in NER, and we manually corrected them as much as possible. We corrected more than 29-K tokens for their POS and NER labels from automatic annotation. From 36,392 TREETAGGER-separated sentence boundaries, we arrived at 42,023 sentences, including those starting without punctuation marks from the previous sentence in the corpus. Because we merged the corpus into a single file to process, the order of sentences was randomized based on TREETAGGER-assigned sentence boundaries using punctuation marks. We split the corpus according to an 80:10:10 ratio for training, development, and testing datasets.

Thereafter, we removed all duplicated sentences. Finally, after splitting, the datasets contained approximately 763-K, 86-K, and 85-K tokens for 33-K, 3-K, and 3-K sentences, respectively. After automatically assigning POS and NER labels using the existing POS and NER models and SBD labels using the heuristic rules, it took longer than 2 weeks (approximately 80 hours) for a language expert to manually verify the labels and correct them. The verification process was straightforward, wherein the expert verified and corrected manually assigned POS, NER, and SBD labels using a simple text editor.

The detailed statistics of the corpus are provided in Table 2. Although the average number of tokens in *all* sentences was 23.42, the average number of tokens in the *middle* of sentences was 52.90, which is much larger. This is mainly because, at the beginning of the TREETAGGER-separated sentence, relatively short noun fragments (e.g., *Les Sables-d'Olonne La Chaume*), such as a place name or a title of the document and paragraph, which are not part of the main sentence, can appear, as seen in Figure 2. Figure 3 shows a preprocessed SBD dataset for training, wherein sent labels were marked and manually verified.

4. SBD as a sequence labeling problem

In this paper, we propose new SBD approaches wherein we consider the task as a sequence labeling problem. We use the first words of sentences and the associated linguistic information to find the beginning. Therefore, the beginning of a sentence is annotated as a label (B-SENT),

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

Table 2. Detailed statistics of the corpus.

	Train	Development	Test	All
Sentences (all)	33,630	3155	3147	39,932
Sentences (*middle*)	4529	548	554	5631
Avg. `B-sent`	1.155	1.209	1.213	1.164
Tokens (all)	763,743	86,816	85,009	935,568
Tokens (*middle*)	239,517	29,112	29,291	297,920

Sentences (*middle*) represent the number of sentences in which a new sentence starts in the middle of the line. No punctuation marks precede these sentences. AVG. B-SENT indicates the average number of `B-sent` between punctuation mark-based separated sentences where they may contain more than one sentence and a fragment. Tokens (*middle*) show the number of tokens in only *middle* sentences where punctuation marks are not preceded. They exclude tokens from (1) punctuation mark-based separated sentences that contain only one sentence and (2) the beginning of the sentence preceded by middle sentences.

Table 3. Summary of SBD as a sequence labeling problem.

CRFs	ROBERTA-SBD	MULTITASK- SBD
Tokens and sequentially predicted POS and NER labels	Vector representations of tokens and gold POS and NER labels	Vector representations of tokens and predicted POS and NER labels

and the capital letter can be used as a feature, such as for the baseline conditional random fields (CRFs) system. Therefore, we need two labels, $\mathscr{Y} = \{\text{B-SENT}, \text{O}\}$ for the current SBD task. To train and evaluate the proposed labeling model, we propose several different models. First, we use CRFs as a baseline labeling algorithm. Second, we implement our own neural baseline for the SBD model using a cross-lingual language model robustly optimized bidirectional encoders from transformers approach (XLM-RoBERTa) (Conneau *et al.* 2020) (ROBERTA-SBD). Third, we propose a multitask model that trains POS, NER, and SBD labels simultaneously (MULTITASK-SBD). Table 3 summarizes the proposed models for SBD as a sequence labeling problem.

4.1 CRF baseline SBD model

An advantage of CRFs, compared with previous sequence algorithms (e.g., Hidden Markov Models (HMMs)), is that we can assign our own defined features. Thus, CRFs can usually outperform HMMs, owing to the relaxation of independence assumptions. We used binary tests for feature functions by distinguishing between unigram ($f_{y,x}$) and bigram ($f_{y',y,x}$) features:

$$f_{y,x}(y_i, x_i) = \mathbf{1}(y_i = y, x_i = x)$$
$$f_{y',y,x}(y_{i-1}, y_i, x_i) = \mathbf{1}(y_{i-1} = y', y_i = y, x_i = x) \quad (1)$$

where $\mathbf{1}(condition) = 1$ if the condition is satisfied and 0 otherwise. ($condition$) represents the input sequence, x, at the current position, i, with CRF label y. We used word, POS, and NER for the input sequence, x, and for the unigram ($w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}$,) and bigram ($w_{t-2}/w_{t-1}, w_{t-1}/w_0, w_0/w_{t+1}, w_{t+1}/w_{t+2}$ for the word, e.g.,) features. In addition to the features, we also used a `Capitalized` feature, where w_t matches `[A-Z][a-z]+`. Previous work on NER utilized text chunking information, which divides a text into phrases in such a way that it syntactically relates words (Tjong Kim Sang and Buchholz 2000). However, we do not do so because detecting phrase boundaries is especially challenging in French, owing to the flat structure of the

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

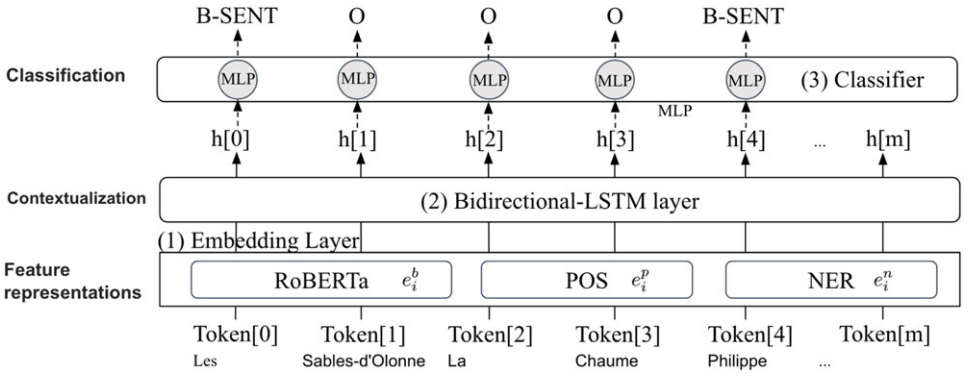

Figure 4. Overall structure of our ROBERTA-SBD model.

French treebank (Abeillé, Clément, and Toussenel 2003), wherein we would otherwise obtain text chunking information. However, it gives ambiguous phrase boundaries.

4.2 Contextualized NN SBD model

We investigated the performance of our system based on the CRF model and checked the effect on the proposed features, including named entities. However, we were still able to explore the performance of SBD using state-of-the-art techniques. In this section, we introduce our own neural network (NN)-based SBD system. Following the previously proposed NN-based systems, we implemented an SBD system using XLM-RoBERTa (Conneau et al. 2020). RoBERTa (Liu *et al.* 2019) is a BERT-like model that applies the masked language model and shows outstanding performance on several NLP tasks (Devlin *et al.* 2018). XLM-RoBERTa is a multilingual RoBERTa trained in several languages. Figure 4 shows the overall structure of our system. The system consumes a list of words $X = (x_1, x_2, ..., x_m)$, where x_i $(1 \leq i \leq m)$ consists of a word form, a POS label, and an NER label. We convert the word form of X to a list of word representations, $E^b = (e_1^b, e_2^b, ..., e_m^b)$ based on the pretrained XLM-RoBERTa model.[b] First, we convert POS and NER to their distributional vector representation by assigning each POS and NER label to randomly initialized embedding vectors, $E^p = (e_1^p, e_2^p, ..., e_m^p)$ and $E^n = (e_1^n, e_2^n, ..., e_m^n)$, respectively. The same value is assigned for the same POS and NER labels. To consider the word form, the POS, and the NER feature as a unified embedding, we concatenate them ($[e_i^b; e_i^p; e_i^n]$) and transform the unified embedding using LSTM as follows:

$$\begin{aligned} e_i &= \left[e_i^b; e_i^p; e_i^n\right] \\ f_i, b_i &= BiLSTM(r_0, (e_1, .., e_m))_i \\ h_i &= \left[f_i; b_i\right] \end{aligned} \tag{2}$$

where r_0 denotes a randomly initialized initial vector for the LSTM hidden layer, f_i is the forward-pass hidden layer of the BiLSTM for word i, b_i is the backward-pass, and h_i is the concatenation of the two. Previous studies have shown that applying LSTM after the concatenation of different embeddings showed better performance because the output of the LSTM keeps track of the contextual information (Lim *et al.* 2018). Finally, we apply a multilayered perceptron (MLP) classifier with a weight parameter, $Q^{(sbd)}$, including a bias term, $b^{(sbd)}$, to classify sentence boundaries for

[b] https://github.com/huggingface/transformers.

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

the output hidden state, h_i, as follows:

$$p_i^{(sbd)} = Q^{(sbd)} MLP(h_i) + b^{(sbd)}$$
$$y_i^{(sbd)} = \underset{j}{\text{argmax}}\, p_{i,j}^{(sbd)} \tag{3}$$

where the value of j in $p_{i,j}^{(sbd)}$ is two because we have two labels (`B-SENT` and `O`) for the sentence boundary. During training, the system adjusts the parameters of the network, θ, that maximizes the probability, $P(Y|X, \theta)$, from the training set, T, based on the conditional negative log-likelihood, sbd-loss(θ). Thus,

$$\text{sbd-loss}(\theta) = \sum_{(X,Y)\in T} -\log P(Y|X, \theta) \tag{4}$$

where $(X, Y) \in T$ denotes an element from the training set, T, a set of sentence boundary labels, Y, and a predicted label of a token, $y_i^{(sbd)}$. We trained our system using a single Adam optimization algorithm (Kingma and Ba 2015) with a cross-entropy loss. During training, we set the number of input batch sizes to 16 and run our system over 100 epochs. For each epoch, we trained our system using only training data and evaluated the validation data. Finally, we selected the best performing model among 50 different ones and run the test data to evaluate the scores.

4.3 Contextualized multitask SBD model

In reality, the input of an SBD task is plain text without any linguistic information. A potential approach may involve a cascade pipeline system, wherein the first tagger assigns POS labels to each word, the second tagger assigns NER labels based on words and POS labels, and the third tagger assigns SBD labels based on previous information. This system uses predicted POS and NER labels incrementally to detect sentence boundaries. However, these tasks can be achieved simultaneously. In this section, we propose a multitask learning scenario that handles POS and NER labeling and SBD labeling simultaneously.

Following the previously investigated multitask learning problem with a shared lexical representation (Hashimoto *et al.* 2016; Lim *et al.* 2018, 2020), we propose a more realistic SBD model that can be deployed as a real-world application. Our method trains POS, NER, and SBD labels simultaneously, rather than applying POS and NER labeling as a separate task. To obtain the predicted POS and NER features, we introduce two different classifiers in the middle of our neural model as follows:

$$p_i^{(pos)} = Q^{(pos)} MLP\left(e_i^b\right) + b^{(pos)}$$
$$y_i^{(pos)} = \underset{k}{\text{argmax}}\, p_{i,k}^{(pos)}$$
$$e_i^p = Embedding^{(pos)}\left(y_i^{(pos)}\right)$$
$$p_i^{(ner)} = Q^{(ner)} MLP\left(e_i^b\right) + b^{(ner)}$$
$$y_i^{(ner)} = \underset{l}{\text{argmax}}\, p_{i,l}^{(ner)}$$
$$e_i^n = Embedding^{(ner)}\left(y_i^{(ner)}\right) \tag{5}$$

where the value of k in $p_{i,k}^{(pos)}$ and l in $p_{i,l}^{(ner)}$ is the number of POS and NER labels, respectively. The $Embedding(pos)$ and $Embedding(ner)$ denote randomly initialized vectors to represent each

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

Table 4. Hyperparameters in neural models.

Component	Value
e_i^b (RoBERTa) dimension	768
e_i^p (POS) dimension	50
e_i^n (NER) dimension	50
Q (parameter) dimension	300
No. BiLSTM layers	2
MLP output dimension	300
MLP activation function	ReLu
Dropout	0.3
Learning rate	0.000005
β_1, β_2 of Adam optimizer	0.9, 0.99
Epoch	100
Batch size	16
Gradient clipping	5.0

POS and NER label. For example, our system computes a logistic value using the *MLP* classifier; thereafter, we can predict a POS label using the *argmax* computation with a logistic value. Finally, the system converts the predicted POS label as a vector representation by the *Embedding* function. During training, we added two additional losses, `pos-loss` and `ner-loss` by changing a set of SBD labels, Y, in (4) to the POS and NER labels, respectively. The multitask loss is defined as follows:

$$\begin{aligned}\text{multi-loss}(\theta) = \alpha\ &\text{sbd-loss} \\ &+ \beta\ \text{pos-loss} + \gamma\ \text{ner-loss}\end{aligned} \tag{6}$$

where each α, β, and γ indicate the ratio of how much the system learns from each task. We empirically set $\{\alpha = 1.5,\ \beta = 0.5,\ \gamma = 1\}$, and the effect of different values for α, β, γ is further discussed in Section 5.2. From a practical perspective, our multitask model has the advantage of producing POS and NER labels alongside SBD labels.

Figure 5 shows the overall structure of our MULTITASK system. The hyperparameter values that we applied for our neural system (ROBERTA-SBD in Section 4.2) and MULTITASK-SBD in Section 4.3) are listed in Table 4.

5. Experiments and results

We use WAPITI as a CRF implementation (Lavergne, Cappé, and Yvon 2010), and our own neural implementations (ROBERTA-SBD and MULTITASK-SBD) for SBD.

5.1 Results

Table 5 shows results of precision, recall, and the F1 score for how we used different linguistic information to improve SBD results using CRFs, ROBERTA-SBD, and MULTITASK-SBD. We also present the SBD results of the *middle*, where sentence boundaries occur in the middle of the sentence without punctuation marks. Overall, each linguistic feature improves the SBD labeling

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

Table 5. SBD results using different linguistic information.

		CRFs by WAPITI		ROBERTA-SBD	MULTITASK- SBD		
		word	+ P+N (p)	word	+P (p)	+N (p)	+ P+N (p)
(all)	P	95.39	97.27	97.35	97.60	**97.66**	97.54
	R	96.84	95.96	98.21	98.28	98.47	**98.60**
	F1 score	96.11	96.61	97.78	97.94	98.06	**98.07**
(middle)	P	73.83	84.48	85.56	86.80	**87.24**	86.75
	R	80.67	78.39	89.89	90.25	91.33	**92.23**
	F1 score	77.10	81.32	87.67	88.49	89.24	**89.41**

POS (+P) and NER (+N) labels or together (+P+N) are used as features alongside word features. We use predicted linguistic labels (+pos and +ner) from the system (p). We show the entire SBD results (all) and those only for sentence boundaries without precedent punctuation marks (middle).

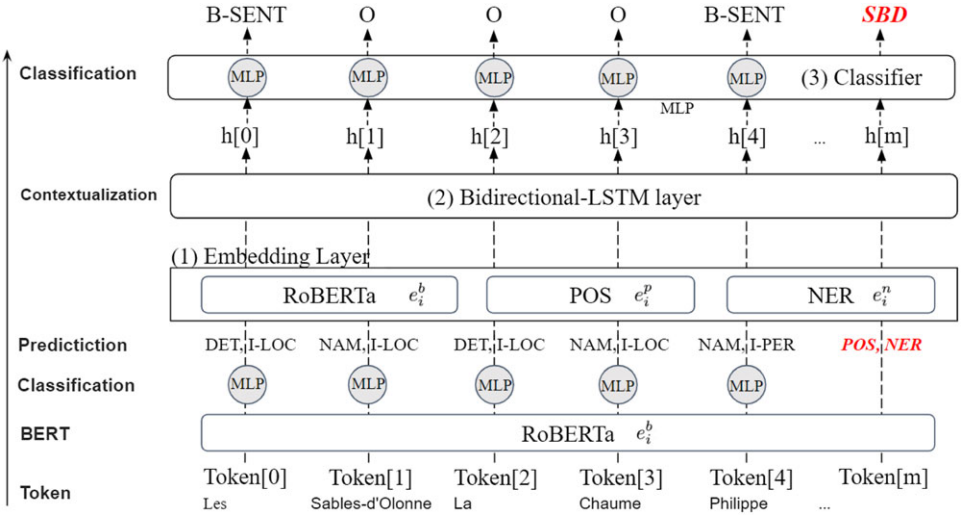

Figure 5. Overall structure of our MULTITASK model.

results for the baseline CRFs and neural models. Although these linguistic features are predicted labels, they can improve SBD results for all experiment settings compared with results that use only word features. Our neural models, ROBERTA-SBD and MULTITASK-SBD, outperformed CRFs for all experimental settings, including word-only (word) and predicted linguistic features (+pos and +ner). We speculate the following plausible explanations. First, the neural model adapts well to the given SBD corpus. Second, the BERT-like model yields more accurate sentence boundaries because it has been trained on a huge number of unlabeled data. Third, the proposed multitask approach efficiently transfers linguistic knowledge by leveraging a shared BERT representation among the following three tasks: POS tagging, NER, and SBD.

5.2 Discussion

5.2.1 Experiments on middle

To clarify our proposal, we show results of SBD without punctuation marks (*middle*) in Table 5, with which existing SBD tools fail to detect sentence boundaries (*middle*), owing to the lack of precedent punctuation marks. For example, when we test PUNKT which is implemented in

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

Table 6. Comparison between the single-task and multitask learning in terms of required computing resources and training time.

	Single-task	Multitask	Gap
GPU (V100)	13.4GB	14.1GB	+0.7GB (5.22%)
Training time	960 seconds	1020 seconds	+60 seconds (6.25%)
Inference time	32 seconds	34 seconds	+2 seconds (6.25%)
Time complexity	$O(n)$	$O(n)$	–

NLTK, it does not have the functionality to detect sentence boundaries without punctuation marks. Moreover, PUNKT obtained a 91.37 F1 score with a precision of 99.80 and a recall of 84.25 on the same SBD evaluation data that we use in Table 5. It is understandable that it obtains high precision because it detects only clear sentence boundaries with punctuation marks. However, its recall is notably low compared with our proposed model because PUNKT is unable to detect sentence boundaries without punctuation marks (*middle*).

5.2.2 POS and NER as features for SBD

We investigated three different scenarios to determine the effect of the proposed features. In Table 5, performances under column WORD, +POS, and +NER denote that the model used the word form only, and either separately with POS labels (+pos) and NER labels (+ner) or accumulatively (+pos+ner). We can see performance improvement when using the predicted POS and NER feature on CRFs, with a gap of +4.22 F1 score in the *middle* scenario. Alternatively, we find a relatively smaller performance gap between the WORD and the +POS+NER models using MULTITASK-SBD with a 1.74 F1 score. Some linguistic features are already captured while training on the unlabeled data based on the masked language modeling for the pretrained BERT model. The WORD as a feature in ROBERTA-SBD already obtains a good result with a 10.57 F1 score compared with the baseline CRFs, and the effect of POS and NER features is relatively smaller than those of the CRF model.

5.2.3 Cost-effectiveness of the proposed system

The multitask model normally requires more computing resources and training time. Therefore, it is important to investigate the cost-effectiveness of the proposed model from a practical point of view. Table 6 shows the comparison between the single and multitasks in terms of cost-effectiveness. The proposed single and multitask models consume 13.4- and 14.1-GB GPU memory, respectively, when training with a batch size of 16. 330-M parameters are required for the single task, which includes in the XLM-RoBERTa model. The single-task model runs for approximately 16 minutes over the training data, and it handles 929 tokens per second. The multitask model runs for 17 minutes and handles 875 tokens. However, it should be noted that the system can yield higher accuracy classifiers at the expense of 6.17% more training time, including predicting POS and NER labels. During the inference phase of the test data, the multitask model can predict 2801 tokens/s for SBD, POS, and NER labels. As our model needs to consider the number of n words as the input, the time complexity is $O(n)$.

6. Ablation study and analysis

6.1 Effect of the multitask learning procedure

As mentioned in Section 4.3, we empirically set the learning weight for each task as $\{\alpha = 1.5, \beta = 0.5, \gamma = 1\}$ for SBD, POS, and NER, respectively. However, performance varies depending

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

Table 7. SBD (middle) results based on different multitask models.

	α	β	γ	SBD	POS	NER
`sequential`				88.69	98.94	94.14
`simultaneous`	0.7	0.1	0.2	88.92	**99.28**	94.16
	1	0	0	87.67	–	–
	0	1	0	–	**99.28**	–
	0	0	1	–	–	**94.55**
	1	1	1	89.01	99.21	94.48
	1.5	0.5	1	**89.41**	99.23	94.54
`+fine-tune`	1.5	0.5	1	89.14	99.15	94.23

The parameter values α, β, and γ are described in (6). In `sequential` POS tagging (accuracy), NER (F1 score), and SBD (F1 score) are sequentially trained.

on the training procedure of the multitask model. Hence, we empirically determined which task would be more significant to SBD performance by using different training procedures and considering the learning weight. We investigated two different learning methods: `sequential` and `simultaneous`. The `sequential` method trains a task only for a certain number of epochs and moves on to train another. In Table 7, `sequential` denotes the performance of the sequential model. Conversely, the `simultaneous` learning method trains three different tasks for every epoch with a particular task's learning weight. Table 7 shows the SBD performance of each learning method. The parameters α, β, and γ indicate the ratio of how much our system learns from each task. For example, the second low, where $\alpha = 1$, $\beta = 0$, and $\gamma = 0$, represents single-task learning for SBD with embedding, $e_i = e_i^b$ in (2). Meanwhile, the seventh row with the parameter values $\{\alpha = 1.5, \beta = 0.5, \text{and } \gamma = 1\}$ represents the application of multitask learning from three different tasks. The `sequential` method trains each task sequentially. We trained a POS tagger for the first 20 epochs and then trained a NER tagger from 20 to 40 epochs. We finally trained SBD from 40 to 80 epochs. Although the three tasks were trained separately, the shared BERT embedding e_i^b was affected by all the tasks by updating the BERT embedding e_i^b. The `sequential` method has the advantage of fine-tuning parameters for the particularities of a single task to bootstrap its final SBD performance. Overall, the `simultaneous` method slightly outperforms the `sequential` method by up to 0.72 in SBD.

However, the single-task learning showed the same or slightly better results than our multitask approach when observing two experimental results that set $\{\alpha = 0, \beta = 1, \text{and } \gamma = 0\}$ for a POS task and $\{\alpha = 0, \beta = 0, \text{and } \gamma = 1\}$ for a NER task, respectively. This is because our multitask model focuses more on the SBD task by learning the POS and NER tasks. By following McCann *et al.* (2018), we also investigated an experiment on the fine-tuning method, which applied both `simultaneous` and `sequential` methods simultaneously. We first trained our model with the `simultaneous` method using parameter values of $\{\alpha = 0, \beta = 1, \text{and } \gamma = 0\}$ until 100 epochs; we then fine-tuned only for SBD for 20 epochs. The result is shown as `+fine-tune`. We found that the fine-tuning method with `simultaneous` does not have positive effects on our model; rather, it shows performance degradation in NER and POS tasks. The main reason for performance degradation might be that the BERT embedding was fully adjusted only for the SBD task during fine-tuning. Thus, POS and NER performances decrease, and the lower-performing POS and NER results affect SBD directly because the system considers the predicted results of the tasks.

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

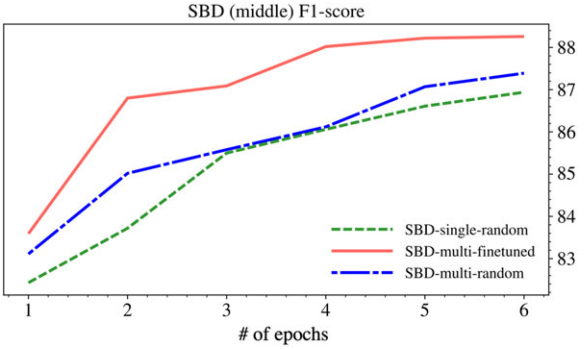

Figure 6. Evaluation results are based on the number of training epochs. The Y-axis represents F1 scores.

6.2 Effect of POS and NER information

Although the proposed multitask learning method has been adapted to several NLP tasks, it has not been explained in detail how multitask learning leverages overall performance. McCann *et al.* (2018) showed an approach to investigating correlations among tasks using multitask learning based on performance changes. The key idea is to test whether a pretrained model trained by a task can leverage a new task's performance. Inspired by the previously proposed method, we investigated the following two questions: "does a pretrained model learn from POS and NER tasks to improve SBD performance?" and "does the pretrained model achieve a better performance than a randomly initialized model in early training?" We assume that if the learned knowledge trained by POS and NER tasks is transferable, it positively affects training SBD during the early epochs of training. In Figure 6, `SBD-single-random` shows that the model trained only for SBD as a single-task learning, `SBD-multi-random` is the model trained on multitask learning with $\{\alpha = 1.5$, $\beta = 0.5$, and $\gamma = 1\}$, and `SBD-multi-fine-tuned` represents the pretrained model trained by POS and NER tasks for 20 epochs, respectively. We observe that the `SBD-multi-fine-tuned` model outperforms the other models over the first six epochs. The average performance gap between `SBD-multi-fine-tuned` and `SBD-single-random` has a 1.78 F1 score during the first six epochs. We conjecture that the BERT embedding in the single model could not obtain any syntactic and named entity information from the POS and NER tasks, whereas the fine-tuned model's BERT acquired general syntactic information from them, and the informed linguistic knowledge was transferred to a new task for SBD. In contrast, as shown in Table 7, the fine-tuned model performed worse than the SBD-multi-random model when training for more than 20 epochs (i.e., `+fine-tune` and `simultaneous`) as described previously, where the BERT embedding was adjusted only for the SBD task.

6.3 Effect of the size of training data

In low-resource NLP, which frequently occurs in real-world settings, the size of training data matters. Figure 7 shows the evaluation results based on `sbd-single` and `sbd-multi` as well as POS and NER. This shows that their results can converge after using 5000 sentences, and `sbd-multi` always outperforms `sbd-single`. In particular, the multitasking setting still performs better than `sbd-single` when `sbd-multi` only utilizes 70% of the training dataset (approximately 20,000 sentences).

6.4 Limitation of the proposed model

The currently proposed system is highly reliant on named entity information in the dataset where results using +NER (+N) and +POS+NER (+P+N) features in MULTITASK-SBD in Table 5

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

Table 8. Multitask SBD (middle) results based on different BERT models.

BERT Model	α	β	γ	SBD	POS	NER
`multilingual-bert`	1.5	0.5	1	87.91	99.18	94.34
`xlm-roberta`	1.5	0.5	1	89.41	**99.23**	**94.54**
`CamemBERT`	1.5	0.5	1	**89.52**	99.16	94.46

The parameter values α, β, and γ are described in (6). POS tagging (accuracy), NER (F1 score), and SBD (F1 score) are presented.

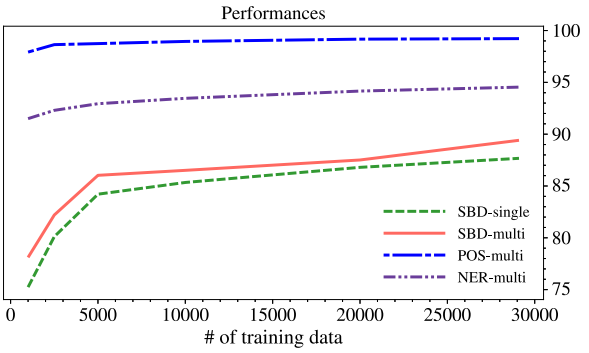

Figure 7. Evaluation results based on the number of training data. The Y-axis represents F1 scores for SBD and NER and accuracy for POS tagging.

are similar. However, improving NER results using extrinsic factors, such as adding additional pseudo datasets by semi-supervised learning, is very difficult, as demonstrated by Park (2018) for French. Although different learning algorithms show various F1 score results ranging from 45.76 (HMM) to 76.26 (bi-LSTM) using the French NER data provided by Europeana Newspapers, semi-supervised learning, in which we automatically annotate a large monolingual corpus, could not improve NER results significantly for F1 scores of 49.69 (HMM) to 76.65 (bi-LSTM). In the proposed multitask learning model, we can still improve the NER results in areas in which further improvement would otherwise be difficult, apart from introducing a completely different and new learning mechanism.

6.5 Comparison between multilingual and French monolingual BERTs

We have shown that the BERT models outperform the CRF model. However, the BERT and the XLM-RoBERTa models that we used are multilingual. There is also a French monolingual BERT proposed by Martin *et al.* (2020) (`CamemBERT`). The monolingual French BERT model was trained only with French plain text from Wikipedia and corpora taken by the Common Crawler. Table 8 shows the ablation study on multilingual and French monolingual BERT models. However, the `CamemBERT` model performs better for SBD. It shows relatively lower-performing results for POS and NER tasks. We leave the detailed discussion on performance between multilingual and French monolingual BERTs as future work.

6.6 Experiments on the heterogeneous domain

The Europarl corpus (Koehn 2005) provides French translations of Europarl proceedings, and it is a good candidate for evaluating our SBD models for general heterogeneous purposes. We

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

Table 9. SBD results of heterogeneous domains using the Europarl corpus.

		CRFs by WAPITI		ROBERTA-SBD	MULTITASK-SBD		
		word	+ P+N (p)	word	+P (p)	+N (p)	+ P+N (p)
(all)	P	99.17	99.18	99.50	99.66	99.53	**99.78**
	R	**99.86**	99.79	99.18	99.19	99.19	99.19
	F1 score	**99.51**	99.48	99.33	99.42	99.36	99.48
(middle)	P	2.59	3.53	8.59	10.32	8.73	**17.32**
	R	**14.10**	12.71	4.47	4.47	5.17	5.17
	F1 score	4.37	5.52	5.88	6.23	6.49	**7.97**

constructed an evaluation dataset from the data of Q4/2000 (October–December 2000) in which the same portion was used for machine translation evaluation. It contained approximately 50-K sentences and 150-K words. To prepare evaluation data of the Europarl corpus, we first detected the beginning sequences of the sentence by using XML tags in which each was also the beginning of a sentence as shown in Figure 8. We assigned POS labels and punctuation mark-based sentence boundaries as described in Section 3.2. We also assigned NER labels as described in Section 3.3. There was an empty line for each punctuation mark-based boundary-detected sentence for rough sentence boundaries. There was no empty line for sentence boundaries only XML tags for evaluation purposes to allow the proposed models to automatically detect them. Although we did not verify automatically assigned POS and NER labels, we manually checked `B-SENT` labels. The dataset contains over 50-K sentences with 1.4 M tokens, with a small ratio (0.0084) of the *middle*. Table 9 shows the results (F1 scores) using CRFs, ROBERTA-SBD and MULTITASK-SBD as experiments in one of the heterogeneous domains. Although overall results were promising, results on *middle* were much lower. As one of the characteristics of the Europarl corpus, as seen in Figure 8, there are noun phrase fragments lacking named entities. The distinction between a noun phrase fragment and the following sentence is immensely challenging in the SBD system, but the semantic property of the noun phrase fragment should be identified. The proposed method using multitask learning relies on sequence-level shallow linguistic information, such as POS and named entities. Even deep linguistic processing, such as syntactic analysis, may not distinguish between the noun phrase fragment and the following sentence because such fragments can be considered as an adverbial phrase in the sentence. Although such a sentence can be considered as being grammatically correct, it is not semantically acceptable. The proposed model attempts to resolve this linguistically difficult problem in SBD by using currently exploitable linguistic properties. Notably, existing SBD methods in previous work cannot detect such boundaries at all, as we showed in Section 1. Additionally, sentence boundaries without punctuation marks, as shown in Figure 8, can be easily remedied by simple heuristics (e.g., using `title` or `subtitle` tags in XML). However, we did not use explicit information to conduct experiments with more realistic conditions, which are not always available in heuristics.

6.7 Experiment on domain adaptation

As was shown in the previous section, the overall performance of our model is promising, although it is still poor in detecting *middle* sentence boundaries in a heterogeneous domain. Generally, a domain adaptation approach would be a good solution to solve such a problem. Table 10 reports on the performance of our model with a domain adaptation method. We split the

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

(a)

```
<CHAPTER ID=1>
Ouverture de la session annuelle
<SPEAKER ID=1 NAME="Le Président">
Je déclare ouverte la session 2000-2001 du Parlement européen.
<CHAPTER ID=2>
Ordre du jour
<SPEAKER ID=2 NAME="Lannoye">
Monsieur le Président, le deuxième point de l'ordre du jour de ce matin est la
recommandation pour la deuxième lecture concernant les produits de cacao et
de chocolat, pour laquelle je suis rapporteur. J'ai appris hier, tout à fait
incidemment, à 20 h 30, que le vote aurait lieu ce midi. ...
```

Europarl corpus for French with XML tags

(b)

```
Ouverture de la session annuelle
Je déclare ouverte la session 2000-2001 du Parlement européen.
Ordre du jour
Monsieur le Président, le deuxième point de l'ordre du jour de ce matin est la
recommandation pour la deuxième lecture concernant les produits de cacao et
de chocolat, pour laquelle je suis rapporteur. J'ai appris hier, tout à fait
incidemment, à 20 h 30, que le vote aurait lieu ce midi. ...
```

Raw data extracted from the Europarl corpus for French

Figure 8. Example of the Europarl corpus for French: (translation) *"Opening of the session I declare resumed the 2000-2001 session of the European Parliament. Agenda Mr President, the second item on this morning's agenda is the recommendation for second reading on cocoa and chocolate products, for which I am the rapporteur. Quite by accident I learnt yesterday, at 8.30 p.m., that the vote was to take place at noon today."* We note that *Je déclare ouverte . . .* ('I declare resumed ...') and *Monsieur le Président, le deuxième ...* ('Mr President, the second ...') are considered as *middle* sentences because punctuation marks are not preceded.

Table 10. SBD results of domain adaptation using the Europarl corpus based on MULTITASK-SBD with +P+N (p).

	(all)			(middle)		
	P	R	F1 score	P	R	F1 score
Out-of-domain	99.58	99.27	99.36	8.32	5.36	6.49
In-domain	99.94	99.69	99.81	89.83	63.09	74.12
Domain adaptation	99.95	99.69	**99.82**	91.37	63.09	**74.64**

Europarl corpus in an 80:10:10 ratio for training, development, and testing datasets, respectively. We performed three different experiments: (1) *out-of-domain*, which determines the performance of our previous model (using the proposed dataset) onto the new Europarl test dataset; (2) *in-domain*, which uses the Europarl dataset both for training and evaluation; (3) *domain adaptation*, which fine-tunes our previous model using the Europarl development dataset to evaluate the Europarl test dataset. Based on the *domain adaptation* approach, we can observe performance improvement in detecting *middle* sentence boundaries by a 0.52 F1 score as well as the performance of *all* where it remains slightly high compared to the *in-domain* model.

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

```
GLANGES (Cramarigeas, Le Châtaignier) Gaston et Marie-Claude, ses enfants ; Laurent
et Christelle, ses petits-enfants ; Evelyne, Guillaume, ses arrière-petits-enfants,
Ainsi que toute la famille et ses amis ont la tristesse de vous faire part du décès
de Madame Lucienne survenu à l'âge de 84 ans. Ses obsèques auront lieu le lundi 19
octobre 2009, à 14h30, en l'église de Glanges. Condoléances sur registre à l'église.
La famille remercie par avance toutes les personnes qui prendront part à sa peine.
PF Graffeuil-Feisthammel, St-Germain-les-Belles.
```

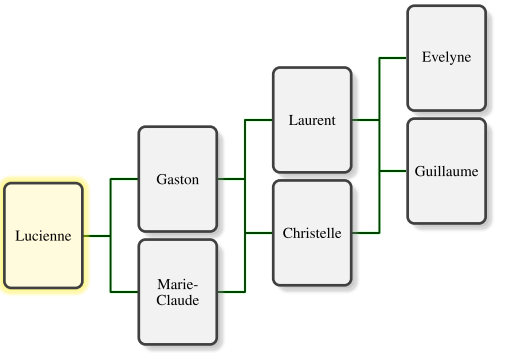

Figure 9. Sentence example for obituaries and possible genealogy tree diagram: (translation) "*GLANGES (Cramarigeas, Le Ch^ataignier) Gaston and Marie-Claude, his children; Laurent and Christelle, his grandchildren; Evelyne, Guillaume, her great-grandchildren, As well as all the family and her friends, are sad to inform you of the death of Madame Lucienne at the age of 84 years. Her funeral will take place on Monday, October 19, 2009, at 2:30 p.m., in the church of Glanges. Condolences on register at the church. The family thanks in advance all the people who will take part in their grief. PF Graffeuil-Feisthammel, St-Germain-les-Belles.*"

6.8 Extrinsic evaluation for SBD

We use a semantic relation recognition task for extrinsic evaluation of SBD. A semantic relation recognition task consists of the automatic recognition of semantic relationships between pairs of entities in a text. Since Roth and Yih (2004) proposed semantic relation recognition data for English, several works on semantic relation recognition have been explored and proposed. This section describes our semantic relation recognition system for French using the proposed SBD system with a dataset for extrinsic evaluation. We specified (1) a segmentation problem to help decide whether a sentence in the obituary contains the information of kinship relations for the deceased person, and (2) a classification problem to decide kinship relations of the deceased person for semantic relation recognition of genealogical relation. The problem of being *survived by whom* identifies which kinship relations in obituaries should be considered for the deceased person after determining whether the information of kinship relations is to appear. Figure 9 provides an example of an obituary, which contains the information of kinship relations and its possible genealogy tree diagram. *Lucienne* (a deceased person) is survived by *Gaston et Marie-Claude* (her children) and other grandchildren and great-grandchildren. An end-to-end semantic relation recognition system uses the heuristic symbolic rules to fill tabular cells based on kinship-related words after detecting sentences in obituaries.

We consider only direct familial relationships of the deceased person in a genealogy tree, including parents, spouse, children, grandchildren, and great-grandchildren. We obtained 3000 additional random obituary documents crawled from the internet to evaluate the end-to-end system and analyze the results. Because the deceased person is given by the document, we identified kinship relations of the deceased person. Table 11 shows the evaluation and statistics of results from the end-to-end system. We noted *misc* relationships for beyond direct relationships, such as (1) siblings (i.e., brother and sister) or other relatives (i.e., second-degree relatives for

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

Table 11. End-to-end system result.

		(total) number
# of deceased persons		2760
a_enfants ('has children')	1534	
a_petits-enfants ('has grandchildren')	984	
a_arrière-petits-enfants ('has great-grandchildren')	362	
a_parents ('has parents')	397	
a_époux ('has a spouse')	816	
misc relationships	2460	
(total) # of relationships		6553

The system extracts 6553 relationships for 2760 deceased persons.

aunt/uncle and niece/nephew), (2) person names without kinship relationships, and (3) other kinship, friend or colleague relationships without person names. 2760 *est_décédé* ('is deceased') relations were given from 3000 documents. 240 missing *est_décédé* were from the original document errors in which they do not explicitly annotate the deceased person. There may be sentences wherein kinship relationships appear without the given *est_décédé* relation. Hence, 1902 sentences where kinship relationships appear were identified using sentence classification from 1878 documents (four false-positive examples). We analyzed 1122 documents where kinship relationships did not appear according to sentence classification results. There were 273 false-negative examples. Although some errors came from results of SBD in which the system cannot provide correct sentence boundaries by merging several sentences because of missing punctuation marks, most presented real classification errors. The number of *a_** ('has ∗') relations presented their occurrences according to the system specification. They also included 545 relations without person names in which only relationship words occurred in the sentence without specific names of kinships. The *misc* relationships contained 415 person names without kinship relationships, and 941 other relationships lacked person names. For evaluation, we manually verified the quality of extracted semantic relations with two native French speakers. By using the proposed SBD system to feed the semantic relation recognition system, the average accuracy for extracted semantic relations was 92.36% by human judgment. When we used the conventional sentence boundary disambiguation system (i.e., TreeTagger) for French, the system did not detect sentence boundaries without full-stop punctuation marks. Therefore, the number of extracted relationships was much smaller because the system could not detect correct sentences that contained information of kinship relations for the deceased person. Additionally, extracted relationships may have been unacceptable because the named entities could not be correctly recognized based on the wrongly segmented sentences.

7. Conclusion

In this paper, we first created a new SBD corpus for French from scratch. We automatically assigned linguistic information and manually corrected them to use the reference corpus. All codes and data are available through author's github.[c] We built our own corpus to measure the SBD result specifically for middle sentences in which a new sentence begins in the middle of

[c]https://github.com/jujbob/frenchSBD.

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

another, and no punctuation marks preceded either one. No previous work has provided such information. Second, we detected the beginning of a sentence without punctuation marks using multitask learning. Sentence boundary disambiguation as a sequence labeling problem is not new (e.g., joint modeling for segmenting tokens and sentences together (Evang et al. 2013; Rei and Søgaard 2018, 2019). However, by introducing linguistic features POS and NER labels, we observed a fair improvement in performance compared to that obtained by features only from the word form. In the ablation study, we demonstrated the effectiveness of the proposed multitask learning procedure and linguistic information. Downstream applications that use SBD results will benefit from the outperformance of the proposed method. Finally, we considered a low-resource NLP setting, which frequently happens in real-world settings, by varying the size of the training data. Even for this scenario, the proposed multitask learning combined with linguistic information outperformed other approaches.

Acknowledgement. The authors would like to thank the reviewers for their insightful suggestions on various aspects of this work. This research was supported by the research fund of Hanbat National University in 2021 for KyungTae Lim.

References

Abeillé A., Clément L. and Toussenel F. (2003). Building a Treebank for French. In Abeillé A. (ed.), *Treebanks: Building and Using Parsed Corpora*. Kluwer, pp. 165–188.

Azzi A.A., Bouamor H. and Ferradans S. (2019). The FinSBD-2019 shared task: Sentence boundary detection in PDF noisy text in the financial domain. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, Macao, China, pp. 74–80.

Björkelund A., Faleńska A., Seeker W. and Kuhn J. (2016). How to train dependency parsers with inexact search for joint sentence boundary detection and parsing of entire documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics, pp. 1924–1934.

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L. and Stoyanov V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 8440–8451.

Dernoncourt F., Lee J.Y. and Szolovits P. (2017). NeuroNER: An easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 97–102.

Devlin J., Chang M., Lee K. and Toutanova K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

Dridan R. and Oepen S. (2013). Document parsing: Towards realistic syntactic analysis. In *Proceedings of The 13th International Conference on Parsing Technologies (IWPT 2013)*, Nara, Japan. Assocation for Computational Linguistics, pp. 127–133.

Evang K., Basile V., Chrupala G. and Bos J. (2013). Elephant: Sequence labeling for word and sentence segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA. Association for Computational Linguistics, pp. 1422–1426.

Gillick D. (2009). Sentence boundary detection and the problem with the US. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, Boulder, Colorado. Association for Computational Linguistics, pp. 241–244.

González-Gallardo C.-E. and Torres-Moreno J.-M. (2017). Sentence boundary detection for French with subword-level information vectors and convolutional neural networks. In *Proceedings of the International Conference on Natural Language, Signal and Speech Processing (ICNLSSP 2017)*, Casablanca, Morocco, pp. 80–84.

González-Gallardo C.-E. and Torres-Moreno J.-M. (2018). WiSeBE: Window-based sentence boundary evaluation. In *Advances in Computational Intelligence: Proceedings of the17th Mexican International Conference on Artificial Intelligence (Part II), MICAI 2018*, Guadalajara, Mexico. Springer International Publishing, pp. 119–131.

Hashimoto K., Xiong C., Tsuruoka Y. and Socher R. (2016). A joint many-task model: Growing a neural network for multiple NLP tasks. CoRR, abs/1611.01587.

Kingma D.P. and Ba J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR 2015)*, San Diego, CA, USA. The International Conference on Learning Representations (ICLR), pp. 1–13.

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

Kiss T. and Strunk J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics* **32**(4), 485–525.

Koehn P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of The Tenth Machine Translation Summit X*, Phuket, Thailand. International Association for Machine Translation, pp. 79–86.

Koehn P., **Hoang H.**, **Birch A.**, **Callison-Burch C.**, **Federico M.**, **Bertoldi N.**, **Cowan B.**, **Shen W.**, **Moran C.**, **Zens R.**, **Dyer C.**, **Bojar O.**, **Constantin A. and Herbst E.** (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic. Association for Computational Linguistics, pp. 177–180.

Lavergne T., **Cappé O. and Yvon F.** (2010). Practical very large scale CRFs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden. Association for Computational Linguistics, pp. 504–513.

Lim K., **Lee J.Y.**, **Carbonell J. and Poibeau T.** (2020). Semi-supervised learning on meta structure: Multi-task tagging and parsing in low-resource scenarios. In *Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, New York, USA. California, USA: AAAI Press, Palo Alto, pp. 8344–8351.

Lim K., **Park C.**, **Lee C. and Poibeau T.** (2018). SEx BiST: A multi-source trainable parser with deep contextualized lexical representations. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium. Association for Computational Linguistics, pp. 143–152.

Liu Y., **Ott M.**, **Goyal N.**, **Du J.**, **Joshi M.**, **Chen D.**, **Levy O.**, **Lewis M.**, **Zettlemoyer L. and Stoyanov**, **V.** (2019). Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692.

Lu W. and Ng H.T. (2010). Better punctuation prediction with dynamic conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA. Association for Computational Linguistics, pp. 177–186.

Manning C.D., **Surdeanu M.**, **Bauer J.**, **Finkel J.R.**, **Bethard S. and McClosky D.** (2014). The stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, Maryland. Association for Computational Linguistics, pp. 55–60.

Martin L., **Muller B.**, **Ortiz Suárez P.J.**, **Dupont Y.**, **Romary L.**, **de la Clergerie R.**, **Seddah D. and Sagot B.** (2020). CamemBERT: A tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 7203–7219.

McCann B., **Keskar N.S.**, **Xiong C. and Socher R.** (2018). The Natural Language Decathlon: Multitask Learning as Question Answering. Technical report, Salesforce Research.

Neudecker C. (2016). An open corpus for named entity recognition in historic newspapers. In Calzolari N., Choukri K., Declerck T., Goggi S., Grobelnik M., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J. and Piperidis S. (eds), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA), pp. 4348–4352.

Palmer D.D. and Hearst M.A. (1997). Adaptive multilingual sentence boundary disambiguation. *Computational Linguistics* **23**(2), 241–267.

Park J. (2018). Le benchmarking de la reconnaissance d'entités nommées pour le français. In *Actes de la Conférence TALN. Volume 1 - Articles Longs, Articles Courts de TALN*, Rennes, France. ATALA, pp. 241–250.

Qi P., **Dozat T.**, **Zhang Y. and Manning C.D.** (2018). Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium. Association for Computational Linguistics, pp. 160–170.

Read J., **Dridan R.**, **Oepen S. and Solberg L.J.** (2012). Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters*, Mumbai, India. The COLING 2012 Organizing Committee, pp. 985–994.

Rei M. and Søgaard A. (2018). Zero-shot sequence labeling: Transferring knowledge from sentences to tokens. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, pp. 293–302.

Rei M. and Søgaard A. (2019). Jointly learning to label sentences and tokens. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, Honolulu, Hawaii, USA, pp. 6916–6923.

Reynar J.C. and Ratnaparkhi A. (1997). A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, DC, USA. Association for Computational Linguistics, pp. 16–19.

Roth D. and Yih W.-t. (2004). A linear programming formulation for global inference in natural language tasks. In Ng H.T. and Riloff E. (eds), *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, Boston, Massachusetts, USA. Association for Computational Linguistics, pp. 1–8.

Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, pp. 252–259.

Tjong Kim Sang E.F. and Buchholz S. (2000). Introduction to the CoNLL-2000 shared task: Chunking. In *Proceedings of the Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop (CoNLL-2000 and LLL-2000)*, Lisbon, Portugal, pp. 127–132.

https://doi.org/10.1017/S1351324922000134 Published online by Cambridge University Press

Treviso M., **Shulby C. and Aluísio S.** (2017). Evaluating word embeddings for sentence boundary detection in speech transcripts. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, Uberlândia, Brazil. Sociedade Brasileira de Computação, pp. 151–160.

Xu C., **Xie L.**, **Huang G.**, **Xiao X.**, **Chng E.S. and Li H.** (2014). A deep neural network approach for sentence boundary detection in broadcast news. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, Singapore. ISCA Archive, pp. 2887–2891.

Cite this article: Lim K and Park J (2024). Real-world sentence boundary detection using multitask learning: A case study on French. *Natural Language Engineering* **30**, 150–170. https://doi.org/10.1017/S1351324922000134