Aligning Opinions: Cross-Lingual Opinion Mining with Dependencies

Mariana S. C. Almeida[*†] **Cláudia Pinto**[*] **Helena Figueira**[*]
Pedro Mendes[*] **André F. T. Martins**[*†]
[*]Priberam Labs, Alameda D. Afonso Henriques, 41, 2º, 1000-123 Lisboa, Portugal
[†]Instituto de Telecomunicações, Instituto Superior Técnico, 1049-001 Lisboa, Portugal
{mla,atm}@priberam.pt

Abstract

We propose a cross-lingual framework for fine-grained opinion mining using bitext projection. The only requirements are a running system in a source language and word-aligned parallel data. Our method projects opinion frames from the source to the target language, and then trains a system on the target language using the automatic annotations. Key to our approach is a novel dependency-based model for opinion mining, which we show, as a byproduct, to be on par with the current state of the art for English, while avoiding the need for integer programming or reranking. In cross-lingual mode (English to Portuguese), our approach compares favorably to a supervised system (with scarce labeled data), and to a delexicalized model trained using universal tags and bilingual word embeddings.

1 Introduction

The goal of **opinion mining** is to extract opinions and sentiments from text (Pang and Lee, 2008; Wilson, 2008; Liu, 2012). With the advent of social media and the increasing amount of data available on the Web, this has become a very active area of research, with applications in summarization of customer reviews (Hu and Liu, 2004; Wu et al., 2011), tracking of newswire and blogs (Ku et al., 2006), question answering (Yu and Hatzivassiloglou, 2003), and text-to-speech synthesis (Alm et al., 2005).

While early work has focused on determining sentiment at document and sentence level (Pang et al., 2002; Turney, 2002; Balog et al., 2006), research has gradually progressed towards **fine-grained opinion mining**, where rather than determining global sentiment, the goal is to parse text into **opinion frames**, identifying opinion expressions, agents, targets, and polarities (Ding et al., 2008), or addressing compositionality (Socher et al., 2013b). Since the release of the MPQA corpus[1] (Wiebe et al., 2005; Wilson, 2008), a standard corpus for fine-grained opinion mining of news documents, a long string of work has been produced (reviewed in §2). Despite the large volume of prior work, opinion mining has by and large been limited to monolingual approaches in English.[2] This is explained by the heavy effort of annotation necessary for current learning-based approaches to succeed, which delays the deployment of opinion miners for new languages.

We bridge the existing gap by proposing a cross-lingual approach to fine-grained opinion mining via **bitext projection**. This technique has been quite effective in several NLP tasks, such as part-of-speech (POS) tagging (Täckström et al., 2013), named entity recognition (Wang and Manning, 2014), syntactic parsing (Yarowsky and Ngai, 2001; Hwa et al., 2005), semantic role labeling (Padó and Lapata, 2009), and coreference resolution (Martins, 2015). Given a corpus of parallel sentences (bitext), the idea is to run a pre-trained system on the source side and then to use word alignments to transfer the produced annotations to the target side, creating an automatic training corpus for the impoverished language.

To alleviate the complexity of the task, we start by introducing a lightweight representation—called **dependency-based opinion mining**—and convert the MPQA corpus to this formalism (§3). We propose a simple arc-factored model that permits easy decoding (§4) and we show that, despite

[1] http://mpqa.cs.pitt.edu/corpora/mpqa_corpus.

[2] Besides English, monolingual systems have also been developed for Chinese and Japanese (Seki et al., 2007), German (Clematide et al., 2012) and Bengali (Das and Bandyopadhyay, 2010).

its simplicity, this model is on par with state-of-the-art opinion mining systems for English (§5). Then, through bitext projection, we transfer these dependency-based opinion frames to Portuguese (our target language), and train a system on the resulting corpus (§6).

As part of this work, a validation corpus in Portuguese with subjectivity annotations was created, along with a translation of the MPQA Subjectivity lexicon of Wilson et al. (2005).[3] Experimental evaluation (§7) shows that our cross-lingual approach surpasses a supervised system trained on a small corpus in the target language, as well as a delexicalized baseline trained using universal POS tags, bilingual word embeddings and a projected lexicon.

2 Related Work

A considerable amount of work on fine-grained opinion mining is based on the MPQA corpus. Kim and Hovy (2006) proposed a method for finding opinion holders and topics, with the aid of a semantic role labeler. Choi et al. (2005) and Breck et al. (2007) used CRFs for finding opinion holders and recognizing opinion expressions, respectively. The two things are predicted jointly by Choi et al. (2006), with integer programming, and Johansson and Moschitti (2010), via reranking. The same method was applied later for joint prediction of opinion expressions and their polarities (Johansson and Moschitti, 2011). The advantage of a joint model was also shown by Choi and Cardie (2010) and Yang and Cardie (2014). Yang and Cardie (2012) classified expressions with a semi-Markov decoder, outperforming a B-I-O tagger; in later work, the same authors proposed an ILP decoder to jointly retrieve opinion expressions, holders, and targets (Yang and Cardie, 2013). A more recent work (İrsoy and Cardie, 2014) proposes a recurrent neural network to identify opinion spans.

All the approaches above rely on a span-based representation of the opinion elements. This makes joint decoding procedures more complicated, since they must forbid overlap of opinion elements or add further constraints, leading to integer programming or reranking strategies. Besides, there is little consensus about what should be the correct span boundaries, the inter-annotator agreement being quite low (Wiebe et al., 2005). In constrast, we use dependencies to model opinion elements and relations, leading to a compact representation that does not depend on spans and which is tractable to decode. A dependency scheme was also used by Wu et al. (2011) for fine-grained opinion mining. Our work differs in which we mine opinions in news articles instead of product reviews, a considerably different task. In addition, the approach of Wu et al. (2011) relies on "span nodes" (instead of head words), requiring solving an ILP followed by an approximate heuristic.

Query-based multilingual opinion mining was addressed in several NTCIR shared tasks (Seki et al., 2007; Seki et al., 2010).[4] However, to our best knowledge, a cross-lingual approach has never been attempted. Some steps were taken by Mihalcea et al. (2007) and Banea et al. (2008), who translated an English lexicon and the MPQA corpus to Romanian and Spanish, but for the much simpler task of sentence-level subjectivity analysis. Cross-lingual sentiment classification was addressed by Wan (2009), Prettenhofer and Stein (2010) and Wei and Pal (2010) at document level, and by Lu et al. (2011) at sentence level. Recently, Gui et al. (2013) applied projection learning for opinion mining in Chinese. However, this work only addresses agent detection and requires translating the MPQA corpus. While all these works are relevant, none addresses fine-grained opinion mining in its full generality, where the goal is to predict full opinion frames.

3 Dependency-Based Opinion Mining

This work addresses various elements of subjectivity annotated in the MPQA corpus, namely:

- direct-subjective expressions (henceforth, **opinions**) that are direct mentions of a private state, *e.g.* opinions, beliefs, emotions, sentiments, speculations, goals, *etc.*;
- the opinion **agent**, *i.e.*, the holder of the opinion;
- the opinion **target**, *i.e.*, what is being argued about;
- the opinion **polarity**, *i.e.*, the sentiment (positive, negative or neutral) towards the target.

As an example, consider the sentence in Figure 1, which has two opinions, expressed by the

[3] The Portuguese corpus and the lexicon are available at `http://labs.priberam.com/Resources`.

[4] NTCIR-8 had a cross-lingual track but in a very different sense: there, queries and documents are in different languages; in contrast, we transfer a model accross languages.

spans "is believed" (O_1) and "are against" (O_2). The first opinion has an implicit agent and a neutral polarity toward the target "the rich elites" (T_1). This target is also the agent (A_2) of the second opinion, which has a negative polarity toward "Hugo Chávez" (T_2).

3.1 Motivation

As noted in prior work (Choi et al., 2005; Kim and Hovy, 2006; Johansson and Moschitti, 2010), one source of difficulty when learning opinion miners on MPQA is with the boundaries of the entity spans. The fact that no criterion for choosing these boundaries is explicitly defined in the annotation guidelines (Wiebe et al., 2005) leads to a low inter-annotator agreement. To circumvent this problem and make the learning task easier, we depart from the classical span-based approaches toward **dependency-based opinion mining**. This decision is inspired by the success of dependency models for syntax and semantics (Buchholz and Marsi, 2006; Surdeanu et al., 2008). These dependency relations can be further converted to opinion spans (as described in §3.3), or directly used as features in downstream applications. As we will see, a compact representation based on dependencies can achieve state-of-the-art results and has the advantage of being easily transferred to other languages through a parallel corpus.

3.2 Dependency Graph

Figure 1 depicts a sentence-level dependency representation for fine-grained opinion mining. The overall structure is a graph whose nodes are head words (plus two special nodes, `root` and `null`), connected by labeled arcs, as outlined below.

Determining head nodes. The three opinion elements that we want to detect (opinions, agents and targets) are each represented by a **head node**, which corresponds to a single word (underlined in Figure 1). When converting the MPQA corpus to dependencies, we determine this "representative" word automatically, by using the following simple heuristic: we first parse the sentence using the Stanford dependency parser (Socher et al., 2013a); then, we pick the last word in the span whose syntactic parent is outside the span (if the span is a syntactic phrase, there is only one word whose parent is outside the span, which is the lexical head). The same heuristic has been used for identifying the heads of mention spans in coreference resolution (Durrett and Klein, 2013).

Defining labeled arcs. The opinion relations are represented as **labeled arcs** that link these head nodes. Two artificial nodes are added: a `root` node, which links to all nodes that represent opinion words, with the label OPINION; and a `null` node, which is used for representing implicit relations. To represent opinion-agent relations, we draw an arc labeled AGENT toward the agent word. For opinion-target relations, the arc is toward the target word and has one of the labels TARGET:0, TARGET:+, or TARGET:-; this encodes the polarity in addition to the type of relation. We also include implicit arcs for opinion elements whose agent or target is not mentioned inside the sentence—these are modeled as arcs pointing to the `null` node.

Dependency opinion graph. We have the following requirements for a well-formed dependency opinion graph:

1. No self-arcs or arcs linking `root` to `null`.
2. An arc is labeled as OPINION if and only if it comes from the `root` node.
3. Arcs labeled as AGENT or TARGET must come from an opinion node (*i.e.*, a node with an incoming OPINION arc).
4. Every opinion node has exactly one AGENT and one TARGET outgoing arcs (possibly implicit).[5]

Similarly to prior work (Choi and Cardie, 2010; Johansson and Moschitti, 2011; Johansson and Moschitti, 2013), we map the MPQA's polarity- into three levels: positive, negative and neutral, where the latter includes spans without polarity annotation or annotated as "both". As in Johansson and Moschitti (2013), we also ignore the "uncertain" aspect of the annotated polarities.

3.3 Dependency-to-Span Conversion

To evaluate the opinion miner against manual annotations and compare with other systems, we need a procedure to convert back from predicted dependencies to spans. In this work, we used a very simple procedure that we next describe,

[5]Even though this assumption is not always met in practice, it is typical in MPQA (only 10% of the opinions have multiple agents, typically coreferent; and only 13% have multiple targets). When multiple agents or targets exist, we keep the ones that are closest to the opinion expression.

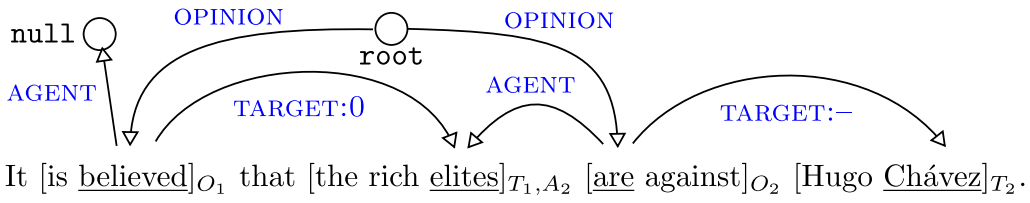

Figure 1: Example of an opinion mining graph in our dependency formalism. Heads are underlined.

which assumes the sentence was previously parsed using a syntactic dependency parser.

To generate agent and target spans, we compute the largest span, containing the head word, whose words are all descendants in the dependency parse tree and that are, simultaneously, not punctuations. To generate opinion spans, we start with the head word and expand the span by adding all neighbouring verbal words. In the case of English, we also allow adverbs, adjectives, modal verbs and the word *to*, when expanding to the left.

The application of this simple approach to the gold dependency graphs in the training partition of the MPQA leads to oracle F_1 scores of 86.0%, 95.8% and 93.0% in the reconstruction of opinion, agent and target spans, respectively, according to the proportional scores described in §5.2.

4 Arc-Factored Model

One of the advantages of the dependency representation is that we can easily decode opinion-agent-target relations without the need of complicated constrained sequence models or integer programming, as done in prior work (Choi et al., 2006; Yang and Cardie, 2012; Yang and Cardie, 2013).

4.1 Decoding

We model dependency-based opinion mining as a structured classification problem. Let x be a sentence and $y \in \mathcal{Y}(x)$ a set of well-formed dependency graphs, according to the constraints stated in §3. We define a score function that decomposes as a sum of labeled arc scores,

$$f(x, y) = \sum_{a \in y} f_a(x, y_a) \quad (1)$$

where y_a is a labeled arc and the sum is over the arcs of the graph y. We use a linear model with weight vector $\boldsymbol{w}$ and local features $\boldsymbol{\phi}_a(x, y_a)$:

$$f_a(x, y_a) = \boldsymbol{w} \cdot \boldsymbol{\phi}_a(x, y_a). \quad (2)$$

For making predictions, we need to compute

$$\widehat{y} = \arg \max_{y \in \mathcal{Y}(x)} f(x, y). \quad (3)$$

Under the assumptions stated in §3, this problem decouples into independent maximization problems (one for each possible opinion word in the sentence). The detailed procedure is as follows, where arcs a can take the form $o \rightarrow h$ (opinion to agent) and $o \rightarrow t$ (opinion to target). For every candidate opinion word o:

1. Obtain the most compatible agent word, $\widehat{h} := \arg\max_h f_{o \rightarrow h}(x, \text{AGENT})$;
2. Obtain the best target word and its polarity, $(\widehat{t}, \widehat{p}) := \arg\max_{t,p} f_{o \rightarrow t}(x, \text{TARGET:}p)$;
3. Compute the total score of this candidate opinion as $s_o := f_{\texttt{root} \rightarrow o}(x, \text{OPINION}) + f_{o \rightarrow \widehat{h}}(x, \text{AGENT}) + f_{o \rightarrow \widehat{t}}(x, \text{TARGET:}\widehat{p})$. Then, if $s_o \geq 0$, add the arcs $\texttt{root} \rightarrow o$, $o \rightarrow \widehat{h}$, and $o \rightarrow \widehat{t}$ to the dependency graph, respectively with labels OPINION, AGENT, and TARGET:$\widehat{p}$.

For a sentence with L words, this decoding procedure takes $O(L^2)$ time. In practice, we speed up this process by pruning from the candidate list arcs whose connected POS were not observed in the training set and whose length were larger than the ones observed in the training set.

4.2 Features

We now describe our features $\boldsymbol{\phi}_a$, which are computed after processing the sentence to predict POS tags, syntactic dependency trees, lemmas and voice (active or passive) information. For English, we used the Stanford dependency parser (Socher et al., 2013a) for the syntactic annotations, the Porter stemmer to compute word stems, and a set of rules for computing the voice of each word. Our Portuguese corpus include all these preprocessing elements (§6.3), with the exception of the voice information (features depending on voice were only used for English).

We also used the Subjectivity Lexicon[6] of Wilson et al. (2005) that we translated to Portuguese

[6] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

(§6.3), and a set of negation words (*e.g. not*, *never*, *nor*) and quantity words (*e.g. very*, *much*, *less*) collected for both languages.

Our arc-factored features are described below; they are inspired by prior work on dependency parsing (Martins et al., 2013) and fine-grained opinion mining (Breck et al., 2007; Johansson and Moschitti, 2013).

Opinion features. We define a set of features that only look at the opinion word; special symbols are used if the opinion is connected to a `root` or `null` node. The features below are also conjoined with the arc label.

- OPINION WORD. The word itself, the lemma, the POS, and the voice. Conjunction of the word with the POS, and of the lemma with the POS.
- BIGRAMS. Bigrams of words and POS corresponding to the opinion word conjoined with its previous (and next) word.
- LEXICON (BASIC). Conjunction of the strength and polarity of the opinion word in the Subjectivity Lexicon[6] (e.g., "weaksubj+neg").
- LEXICON (COUNT). Number of subjective words (total, positive and negative) in a sentence, with and without being conjoined with the polarity of the opinion word in the lexicon.
- LEXICON (CONTEXT). For each word that is in the lexicon and within the 4-word context of the opinion, the form and the polarity of that word in the lexicon, with and without being conjoined with the form and the polarity in the lexicon of the opinion word. Besides the 4-word context, we also used the next/previous word in the sentence which is in the lexicon.
- NEGATION AND QUANTITY WORDS. Within the 4-word context, features indicating if a word is a negation or quantity word, conjoined with the word itself and the opinion word.
- SYNTACTIC PATH. The number of words up to the top of the syntactic dependency tree, and the sequence of POS tags in that path.

Opinion-Argument features. In case of arcs that neither connect to `null` nor `root`, the features above are also conjoined with the binned distance between the two words.For these arcs, we did not use the LEXICON (COUNT)/(CONTEXT) features, but we added features regarding the pair of opinion-argument words (below).

- OPINION-ARGUMENT WORD PAIR. Several conjunctions of word form, POS, voice and syntactic dependency relations corresponding to the pair opinion-argument.
- OPINION-ARGUMENT SYNTACTIC PATH. The syntactic path from the opinion word to the argument, conjoined with the POS and the dependency relations in the path (in Figure 1, for the agent "elites" headed by "are" with relation `nsuj`, we have: "VBP↓NNS" and "`nsuj`↓").

For arcs that neither connect to `null` or `root`, we conjoin voice features with the label, distance, and the direction of the arc. For these arcs, we also include back-off features where the polarity information is removed from the (target) labels.

5 English Monolingual Experiments

In a first set of experiments, we evaluated the performance of our dependency-based model for opinion mining (§3) in the MPQA English corpus.

5.1 Learning

We trained arc-factored models by running 25 epochs of max-loss MIRA (Crammer et al., 2006). Our cost function takes into account mismatches between predicted and gold dependencies, with a cost C_P on labeled arcs incorrectly predicted (false positives) and a cost $C_R = 1 - C_P$ on missed gold labeled arcs (false negatives). The cost C_P, the regularization constant, and the number of epochs were tuned in the development set.

5.2 Evaluation Metrics

Opinion spans (Op.) are evaluated with F_1 scores, according to two matching criteria commonly used in the literature: **overlap matching** (OM), where a predicted span is counted as correct if it overlaps a gold one, and **proportional matching** (PM), proposed by Johansson and Moschitti (2010). For the latter, we use the following formula for the recall, where we consider the sets of gold ($\mathcal{G}$) and predicted ($\mathcal{P}$) opinion spans:[7]

$$R(\mathcal{G}, \mathcal{P}) = \sum_{p \in \mathcal{P}} \max_{g \in \mathcal{G}} \frac{|g \bigcap p|/|p|}{|\mathcal{P}|}; \qquad (4)$$

[7] This metric is slightly different from the PM metric of Johansson and Moschitti (2010), in which recall was computed as $R(\mathcal{G}, \mathcal{P}) = \sum_{p \in \mathcal{P}} \sum_{g \in \mathcal{G}} \frac{|g \cap p|/|p|}{|\mathcal{P}|}$. The reason why we replace the "sum" by a "max" is that each predicted span p in (4) could contribute to the recall with a value greater than 1. Since most of the predicted spans only overlap a single gold span, this fix has a very small effect in the final scores.

the precision is $P(\mathcal{G}, \mathcal{P}) = R(\mathcal{P}, \mathcal{G})$. We also report metrics based on a **head matching** (HM) criterion, where a predicted span is considered correct if its syntactic head matches the head of the gold span. We consider that a pair opinion-agent (Op-Ag.) or opinion-target (Op-Tg.) is correctly extracted according to the OM or the HM criteria, if both the elements satisfy these criteria and the relation holds in the gold data. We also compute the metric described in Johansson and Moschitti (2010) which measures how well agents of opinions are predicted based on a proportional matching (PM) criterion. This metric is applied to evaluate the extraction of both agents and targets. Finally, to evaluate the opinions' polarities (Op-Pol. metric) we consider as correct opinions where the span and polarity both match the gold ones.

5.3 Results: Dependency-Based Model

We assess the quality of our monolingual dependency-based model by comparing it to the recent state-of-the-art approach of Johansson and Moschitti (2013), whose code is available online.[8] That paper reports the performance of a basic span-based pipeline system (which extracts opinions with a CRF, followed by two separate classifiers to detect polarities and agents), and of a more sophisticated system that applies a reranking procedure to account for more complex features that consider interactions accross opinion elements.

We ran experiments using the same data and MPQA partitions as Johansson and Moschitti (2013). However, since our system is designed for predicting opinion, agents and targets together, we removed the documents that were not annotated with targets. The final train/development/test sets have a total of 6,774/1,404/2,559 sentences and 3,834/881/1,426 opinions, respectively.

Table 1 reports the results; since the systems of Johansson and Moschitti (2013) do not predict targets, Table 1 omits target scores.[9] We observe that our dependency-based system achieves results competitive with the best results of Johansson and Moschitti (2013) and clearly above the ones reached by their basic system that does not use re-ranking features. Though the two systems are not fully comparable,[10] the results in Table 1 show that our dependency-based approach (§3.2) followed by a simple dependency-to-span conversion (§3.3) is, despite its simplicity, on par with a top-performing opinion mining system. We conjecture that this is due to the ability to extract opinions, agents, and targets jointly using exact decoding. Note that our proposed dependency scheme would also be able to include additional global features relating pairs of opinions (by adding scores to pairs of opinion arcs) or two opinions having the same agent (by adding scores to pairs of agent arcs sharing its argument), similar to the reranking features used by Johansson and Moschitti (2013). Similar second-order scores have been used in syntactic and semantic dependency parsing (Martins et al., 2013; Martins and Almeida, 2014), but with an increase in the complexity of the model and of the decoder.

6 Cross-Lingual Opinion Mining

We now turn to the problem of learning a opinion mining system for a resource-poor language (Portuguese), in a cross-lingual manner. We use a bitext projection approach (§6.1), whose only requirements are a model for a resource-rich language (English) and parallel data (§6.2).

6.1 Bitext Projection

Our methodology is outlined as Algorithm 1. For simplicity, we call the source and target languages English (e) and "foreign" (f), respectively. The procedure is inspired by the idea of bitext projection (Yarowsky and Ngai, 2001). We start by training an English system on the labeled data $\mathcal{L}^e$ (line 1), which in our case is the MPQA v.2.0 corpus. This system is then used to label the English side of the parallel data, automatically identifying opinion frames (line 2). The next step is to run a word aligner on the parallel data (line 3). The automatic alignments are then used to project the opinion frames to the target language (along with some filtering), yielding an automatic corpus $\widehat{D}^{(f)}$ (line 4), which finally serves to train a system for the target language (line 5).

6.2 Parallel Data

We use an English-Portuguese parallel corpus based on the scientific news Brazilian magazine *Revista Pesquisa FAPESP*, collected by Aziz and

[8] http://demo.spraakdata.gu.se/richard/unitn_opinion/details.html

[9] We will report target scores later in §7.

[10] Our system makes use of target annotations to predict the opinion frames, while Johansson and Moschitti (2013) has access not only to *direct subjective* spans but also to *subjective expressions* annotations with their agents and polarity information.

	JM13, BASIC			JM13, RERANKING			OUR SYSTEM		
	HM	PM	OM	HM	PM	OM	HM	PM	OM
Op.	56.3	56.2	60.6	58.6	59.2	63.7	**61.6***	**59.8**	**65.1**
Op-Ag.	40.3	47.1	44.9	42.4	**51.4**	48.1	**45.7***	**51.4**	**50.3***
Op-Tg.	-	-	-	-	-	-	**31.3***	**48.3***	**48.3***
Op-Pol.	46.1	45.9	49.3	**48.5**	**48.9***	**52.5**	47.9	47.0	50.7

Table 1: Method comparison: F_1 scores obtained in the MPQA corpus, for our dependency based method and the approaches in Johansson and Moschitti (2013), with and without reranking. The symbol * indicates that the best system beats the other systems with statistical significance, with $p < 0.05$ and according to a bootstrap resampling test (Koehn, 2004).

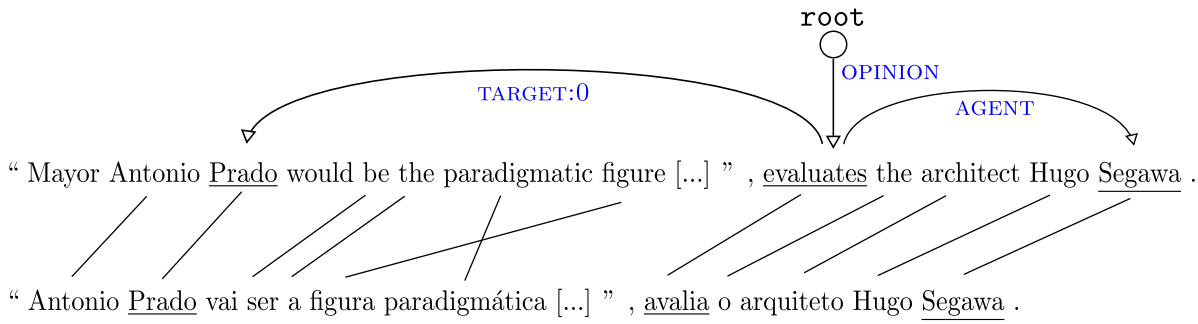

Figure 2: Excerpt of a bitext document from FAPESP, with automatic opinion dependencies. The annotations are directly projected to Portuguese via automatic word alignments.

Algorithm 1 Cross-Lingual Opinion Mining

Input: Labeled data $\mathcal{L}^e$, parallel data $\mathcal{D}^e$ and $\mathcal{D}^f$.
Output: Target opinion mining system $\mathcal{S}^f$.

1: $\mathcal{S}^e \leftarrow \text{LEARNOPINIONMINER}(\mathcal{L}^e)$
2: $\widehat{\mathcal{D}}^e \leftarrow \text{RUNOPINIONMINER}(\mathcal{S}^e, \mathcal{D}^e)$
3: $\mathcal{D}^{e\leftrightarrow f} \leftarrow \text{RUNWORDALIGNER}(\mathcal{D}^e, \mathcal{D}^f)$
4: $\widehat{\mathcal{D}}^f \leftarrow \text{PROJECTANDFILTER}(\mathcal{D}^{e\leftrightarrow f}, \widehat{\mathcal{D}}^e)$
5: $\mathcal{S}^f \leftarrow \text{LEARNOPINIONMINER}(\widehat{\mathcal{D}}^f)$

Specia (2011). Though this corpus is in Brazilian Portuguese (while our validation corpus is in European Portuguese), we preferred FAPESP over other commonly used parallel corpora (such as the Europarl and UN datasets), since it is closer to our newswire target domain, with a smaller prominence of direct speech. We computed word alignments using the Berkeley aligner (Liang et al., 2006), intersected them and filtered out all the alignments whose confidence is below 0.95.

After annotating the English side of FAPESP with the pre-trained system ($\widehat{\mathcal{D}}^e$ in Algorithm 1, with a total of 166,719 sentences and 81,492 opinions), the high confidence alignments ($D^{e\leftrightarrow f}$) are used to project the annotations to the Portuguese side of the corpus. The automatic annotations produced by our dependency-based system are easily transferred at a word level (for words with high confidence alignments), as illustrated in Figure 2. To improve the quality of the resulting corpus, we excluded sentences whose alignments cover less than 70% of the words in the target side of the corpus, or sentences whose opinion elements were not fully projected through high confidence alignments. At this point, we obtain an automatically annotated corpus in Portuguese ($\widehat{\mathcal{D}}^f$), with 106,064 sentences and 32,817 opinions.

6.3 Portuguese Opinion Mining Corpus

For validation purposes, we also created a Portuguese corpus with manually annotated fine-grained opinions. The corpus consists of a subset of the documents of the Priberam Compressive Summarization Corpus[11] (Almeida et al., 2014), which contains 80 news topics with 10 documents each, collected from several Portuguese newspapers, TV and radio websites in the biennia 2010–2011 and 2012–2013. In the scope of the current work, we selected and annotated one document of each of the 80 topics. The first biennium was selected as the test set and the second biennium was split into development and training sets (see Ta-

[11] http://labs.priberam.com/Resources/PCSC

ble 2 for statistics).

	#doc.	#sent.	#opin.
Train	20	441	240
Dev	20	225	197
Test	40	560	391

Table 2: Number of documents, sentences and opinions in the Portuguese Corpus.

	HM	PM	OM
Op.	77.0	76.7	79.2
Op-Ag.	69.1	72.3	73.5
Op-Tg.	61.9	65.4	71.4
Op-Pol.	49.4	49.1	50.7

Table 3: Inter-annotator agreement in the test partition (shown are F_1 scores).

The corpus was annotated in a similar vein as the MPQA (Wiebe et al., 2005), with the addition of the head node for each element of the opinion frame. It includes spans for direct-subjective expressions with intensity and polarity information; agent spans; and target spans. The annotation was carried out by three linguists, after reading the MPQA annotation guidelines (Wiebe et al., 2005; Wilson, 2008) and having a small practice period using the provided examples and some MPQA annotated sentences. Each document was annotated by two of the three linguists and then revised by the third linguist, who (in case of any doubts) discussed with the initial annotators to reach for the final consensus. Scores for inter-annotator agreement are shown in Table 3.

The corpus was annotated with automatic POS tags and dependency parse trees using TurboParser (Martins et al., 2013).[12] We used an in-house lemmatizer to obtain lemmas for each inflected word in the corpus. A Portuguese lexicon of subjectivity was created by translating the words in the Subjectivity Lexicon of Wilson et al. (2005). The annotated corpus and the translated subjectivity lexicon are available at `http://labs.priberam.com/Resources/Fine-Grained-Opinion-Corpus`, and `http://labs.priberam.com/Resources/Subjectivity-Lexicon-PT`, respectively.

[12] `http://www.ark.cs.cmu.edu/TurboParser`

	OUR SYSTEM			DELEXICALIZED		
	HM	PM	OM	HM	PM	OM
Op.	**65.7**	**63.5**	**69.8**	50.1	45.8	52.7
Op-Ag.	**47.6**	**48.8**	**51.1**	33.8	34.8	35.7
Op-Tg.	**34.9**	**44.8**	**50.3**	19.9	28.0	32.1
Op-Pol.	**51.5**	**50.2**	**54.4**	36.7	34.7	38.8

Table 4: F_1 scores obtained in English (MPQA), for our full system and the DELEXICALIZED one.

7 Cross-Lingual Experiments

In a final set of experiments, we compare three systems of fine-grained opinion mining for Portuguese. All were trained as described in §5.1.

7.1 System Description

Baseline #1: Supervised System. A SUPERVISED system was trained on the small Portuguese training set described in §6.3. Though being a small training corpus, this is, to the best of our knowledge, the only existing corpus with fine-grained opinions in Portuguese. We used the same arc-factored model and features described in §4.

Baseline #2: Delexicalized System with Bilingual Embeddings. This baseline consists of a direct model transfer: a DELEXICALIZED system is trained in the source language, without language specific features, so that it can be directly applied to the target language. Despite its simplicity, this strategy managed to provide a fairly strong baseline in several NLP tasks (Zeman and Resnik, 2008; McDonald et al., 2011; Søgaard, 2011).

To achieve a unified feature representation, we mapped all language-specific POS tags to universal tags (Petrov et al., 2012), and removed all features depending on the dependency relations, but maintained those depending on the syntactic path (but not on the dependency relations themselves). In addition, we replaced the lexical features by 128-dimensional cross-lingual word embeddings.[13] To obtain these bilingual neural embeddings, we ran the method of Hermann and Blunsom (2014) on the parallel data (§6.1). We scaled the embeddings by a factor of 2.0 (selected on the dev-set), following the procedure described in Turian et al. (2010).

We trained the English delexicalized system on the MPQA corpus, using the same test documents

[13] A delexicalized system trained without the word embeddings had a worse performance.

	BASELINE #1 (SUP.)			BASELINE #2 (DELEX.)			BITEXT PROJECTION		
	HM	PM	OM	HM	PM	OM	HM	PM	OM
Op.	49.4	48.7	50.8	33.1	32.1	34.3	**58.0***	**55.7***	**58.0***
Op-Ag.	23.5	27.2	31.5	14.3	18.8	20.0	**30.8***	**31.2***	**36.2***
Op-Tg.	23.0	24.9	30.6	11.0	15.7	19.0	**29.4***	**29.4***	**35.6***
Op-Pol.	24.1	23.8	24.7	16.6	16.4	17.6	**35.7***	**34.1***	**35.7***

Table 5: Comparison of cross-lingual approaches. F_1 scores obtained in our Portuguese validation corpus using: a SUPERVISED system trained on the small available data, a DELEXICALIZED system trained with universal POS tags and multilingual embeddings and our BITEXT PROJECTION OF DEPENDENCIES. The symbol * indicates that the best system beats the other systems with statistical significance, with $p < 0.05$ and according to a bootstrap resampling test (Koehn, 2004).

as Riloff and Wiebe (2003) and whose list is available with the corpus, but selecting only documents annotated with targets. We randomly split the remaining documents into train and development sets, respectively with a total of 6,471 and 782 sentences.[14] Table 4 shows the performance of the delexicalized baseline in English, compared with a lexicalized system. We will see how this model behaves in a cross-lingual setting in §7.2.

Our System: Bitext Projection of Opinion Dependencies. Finally, we implemented our cross-lingual BITEXT approach (§6). We trained the (lexicalized) English model on the MPQA corpus (the performance of this model is shown in Table 4). Then, we ran this model on the English side of the parallel corpus, generating automatic annotations, and projected these annotations to the Portuguese side, as described in §6.2. Finally, a Portuguese model was trained on these projected annotations using the arc-factored model and features described in §4.

7.2 Comparison

Table 5 shows the F_1 scores obtained by the three systems on the Portuguese test partition. We observe that the BITEXT approach outperformed the SUPERVISED and the DELEXICALIZED ones in all metrics with a considerable margin, which shows the effectiveness of our proposed method. The SUPERVISED system suffers from the fact that the training set is too small to allow good generalization; the bitext projection method, in contrast, can create arbitrarily large training corpora without any annotation effort. The performance of the DELEXICALIZED system is rather disappointing. This result is justified by a decrease of performance in English due to the delexicalization (cf. Table 4), followed by an extra loss of quality due to language differences.

Though our BITEXT approach scores the best, the scores are behind the range of values obtained for English (Table 4), and far from the inter-annotator agreement numbers (Table 3), suggesting room for improvement. The polarity scores in Table 5 appear to be relatively low. This fact is probably be justified with the annotator agreement scores (Table 3) which are considerably lower for these metrics.

8 Conclusions

We presented a cross-lingual framework for fine-grained opinion mining. We used a bitext projection technique to transfer dependency-based opinion frames from English to Portuguese. Experimentally, our dependency model achieved state-of-the-art results for English, and the Portuguese system trained with bitext projection outperformed two baselines: a supervised system trained on a small dataset, and a delexicalized model with bilingual word embeddings.

9 Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments, and Richard Johansson for sharing his code and for answering several questions. This work was partially supported by the EU/FEDER programme, QREN/POR Lisboa (Portugal), under the Intelligo project (contract 2012/24803) and by a FCT grants UID/EEA/50008/2013 and PTDC/EEI-SII/2312/2012.

[14] Note that this split is different from the one we used in §5. There we used the same split as Johansson and Moschitti (2013), for a fair comparison with their system; here, we follow the standard MPQA test partition.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *HLT and EMNLP*.

Miguel B Almeida, Mariana SC Almeida, André FT Martins, Helena Figueira, Pedro Mendes, and Cláudia Pinto. 2014. Priberam compressive summarization corpus: A new multi-document summarization corpus for european portuguese. In *LREC*.

Wilker Aziz and Lucia Specia. 2011. Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In *STIL 2011*.

Krisztian Balog, Gilad Mishne, and Maarten de Rijke. 2006. Why are they excited?: Identifying and explaining spikes in blog mood levels. In *EACL*.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *EMNLP*.

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *IJCAI*, pages 2683–2688.

S. Buchholz and E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Int. Conf. on Natural Language Learning*.

Yejin Choi and Claire Cardie. 2010. Hierarchical sequential learning for extracting opinions and their attributes. In *ACL*, pages 269–274.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proc. of HLT and EMNLP*.

Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *EMNLP*.

Simon Clematide, Stefan Gindl, Manfred Klenner, Stefanos Petrakis, Robert Remus, Josef Ruppenhofer, Ulli Waltinger, and Michael Wiegand. 2012. MLSA A Multilayered Reference Corpus for German Sentiment Analysis. In *LREC*, Istanbul, Turkey, may.

K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. 2006. Online Passive-Aggressive Algorithms. *Journal of Machine Learning Research*, 7:551–585.

Dipankar Das and Sivaji Bandyopadhyay. 2010. Labeling emotion in bengali blog corpus a fine grained tagging at sentence level. In *In Proc. of the 8th Workshop on Asian Language Resources (ALR8), COLING-2010*, pages 47–55.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *WSDM*, pages 231–240.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *EMNLP*.

Lin Gui, Ruifeng Xu, Jun Xu, and Chenxiang Liu. 2013. A cross-lingual approach for opinion holder extraction. *Journal of Computational Information Systems*, 9(6):2193–2200.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributional Semantics. In *ACL*.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *AAAI*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Ozan İrsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. In *EMNLP*, pages 720–728, Doha, Qatar, October.

Richard Johansson and Alessandro Moschitti. 2010. Reranking models in fine-grained opinion analysis. In *COLING*.

Richard Johansson and Alessandro Moschitti. 2011. Extracting opinion expressions and their polarities: exploration of pipelines and joint models. In *ACL:HLT*.

Richard Johansson and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3):473–509.

Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *SST*.

P. Koehn. 2004. Statistical signicance tests for machine translation evaluation. In *ACL*.

Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *NAACL*.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K. Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *ACL:HLT*.

André F. T. Martins and M. S. C. Almeida. 2014. Priberam: A turbo semantic parser with second order features. In *SemEval*, pages 471–476, August.

André F. T. Martins, Miguel B. Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *ACL*.

André F. T. Martins. 2015. Transferring coreference resolvers with posterior regularization. In *ACL*.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *EMNLP*, pages 62–72.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *ACL*.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *EMNLP*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *LREC*.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *ACL*.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *EMNLP*.

Yohei Seki, David Kirk Evans, Lun-Wei Ku, Hsin-Hsi Chen, Noriko Kando, and Chin-Yew Lin. 2007. Overview of opinion analysis pilot task at NTCIR-6. In *NTCIR-6 Workshop Meeting*.

Yohei Seki, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, and Noriko Kando. 2010. Overview of opinion analysis pilot task at NTCIR-8: A Step Toward Cross Lingual Opinion Analysis. In *NTCIR-8 Workshop Meeting*.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013a. Parsing with compositional vector grammars. In *ACL*.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pages 1631–1642.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *ACL:HLT short papers*, pages 682–686. Association for Computational Linguistics.

M. Surdeanu, R. Johansson, A. Meyers, L. Màrquez, and J. Nivre. 2008. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. In *CoNLL*.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Trans. of the Association for Computational Linguistics (TACL)*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *ACL*.

Peter D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *ACL*.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *ACL-IJCNLP*.

Mengqiu Wang and Chris Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. *TACL*, 2:55–66.

Bin Wei and Christopher Pal. 2010. Cross lingual adaptation: an experiment on sentiment classifications. In *ACL*, pages 258–262.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT and EMNLP*.

Theresa Wilson. 2008. *Fine-Grained Subjectivity Analysis*. Ph.D. thesis, University of Pittsburgh.

Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu. 2011. Structural opinion mining for graph-based sentiment representation. In *EMNLP*.

Bishan Yang and Claire Cardie. 2012. Extracting opinion expressions with semi-markov conditional random fields. In *EMNLP-CoNLL*.

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *ACL*.

Bishan Yang and Claire Cardie. 2014. Joint modeling of opinion expression extraction and attribute classification. *Trans. of the Association for Computational Linguistics (TACL)*.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *NAACL*.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *EMNLP*.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP*, pages 35–42.