Determining Event Durations: Models and Error Analysis

Alakananda Vempala, Eduardo Blanco and Alexis Palmer
University Of North Texas
alakanandavempala@my.unt.edu, eduardo.blanco@unt.edu
alexis.palmer@unt.edu

Abstract

This paper presents models to predict event durations. We introduce aspectual features that capture deeper linguistic information than previous work, and experiment with neural networks. Our analysis shows that tense, aspect and temporal structure of the clause provide useful clues, and that an LSTM ensemble captures relevant context around the event.

1 Introduction

Robust textual understanding requires identifying events and temporal relations between them. Beyond event participants, a crucial piece of information regarding events is their duration, an attribute rarely mentioned explicitly. For example, *taking a shower* lasts a few minutes (not days), and a *vacation* lasts a few days (not years). Core tasks such as temporal understanding and reasoning, as well as applications such as temporal question answering (Llorens et al., 2015) would benefit from knowing the expected duration of events.

Consider a system that extracts temporal relations such as IS_INCLUDED (Cassidy et al., 2014, among others). When deciding whether a relation holds between an event and a temporal expression, such a system would benefit from knowing the duration of the event at hand. For example, argument *y* of IS_INCLUDED(*built a house*, *y*) must be a temporal span ranging from a few weeks to a year—the expected duration of *built a house*. Thus relation candidates such as IS_INCLUDED(*built a house*, *4/5/2016*) could be discarded right away.

Similarly, event durations combined with event ordering and temporal anchoring would help to determine the time of subsequent events. For example, if John Doe started *his drive to work* at 8:00am, it is reasonable to expect him to start working by 9:00am because commuting took him (most likely) between a few minutes to an hour.

In this paper, we classify events based on their expected duration. Specifically, we differentiate between events whose duration is less than a day, and events whose duration is a day or more. The main contributions are: (a) linguistically motivated features that yield better results than previous work, (b) an LSTM ensemble that obtains the best results to date, and (c) error analysis shedding light on the benefits of our models.

2 Related Work

TimeBank (Pustejovsky et al., 2006) is the corpus of reference for temporal information. The annotations follow TimeML (Pustejovsky et al., 2010) and include events, temporal expressions (e.g., *last Friday*), temporal signals (e.g., *when*, *during*), and links encoding relations. TimeBank does not annotate the expected duration of events.

Annotating and learning event durations was pioneered by Pan et al. (2011), who annotated the events in TimeBank with their expected durations. Gusev et al. (2011) use query patterns in an unsupervised approach to predict the duration of events. The work presented here builds upon these previous works: we introduce additional features and an LSTM ensemble that obtains the best results to date. The new features are inspired by previous work on assigning situation entity (SE) type labels to clauses (Friedrich et al., 2016). SE types are a linguistic categorization of semantic clause type, whereby each clause is labeled according to the type of situation it introduces to a discourse (STATE, EVENT, GENERIC, and GENERALIZING SENTENCE (also known as habituals)).

Other related works include efforts modeling event durations in social media (Williams and Katz, 2012), and temporal anchoring of, among others, durative events (Reimers et al., 2016).

Proceedings of NAACL-HLT 2018, pages 164–168
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

	#	Description
Pan et al.	1-3	event token, lemma and POS tag
	4-9	head word, lemma and POS tags of the syntactic subject and object of the event
	10-18	three closest hypernyms of the event, subject and object
Gusev et al.	19-20	named entity types of the syntactic subject and object of the event
	21	flag indicating if the event is a reporting verb
	22-25	flags indicating presence of *dobj*, *iobj*, *pobj* and *advmod* syntactic dependencies of the verb
Aspectual features inspired by situation entities (Friedrich et al.)	26	event tense: past, present or future, and simple, perfect or continuous form
	27	whether the event is in active or passive voice
	28	type of determiner present in the subject
	29	noun type of the subject
	30	subject person
	31	whether the subject is a bare plural
	32-40	synset id of the two closest hypernyms in WordNet of the event, subject and object
	41-43	lexical filename of the event, subject and object in WordNet
	44-46	depth of the event, subject and object in the WordNet taxonomy
	47	countability from WebCelex of the subject and object
	48	number of modifiers in the sentence
	49	adverbial degree of the sentence
	50	whether the sentence contains an adverb
	51-700	flags indicating the Brown clusters present in the sentence

Table 1: Feature set to predict the expected duration of events with SVM. Features 1–25 were previously proposed for the same task. Features 26–700 are inspired by previous work assigning situation entity types to clauses (2016).

3 Corpus

We use the corpus by Pan et al. (2011), who annotated the events in TimeBank (Pustejovsky et al., 2003) with their expected durations by specifying upper and lower bounds. The authors clustered these bounds into two labels: less than a day (<day) and a day or longer (≥day), and the corpus contains 2,354 events (<day: 958, ≥day: 1,396). The same event predicate may have different durations depending on context as exemplified below:

- *I* <u>*want*</u> *to be absolutely clear, to the extent there is any implication that Mrs. Currie believes that the President or anyone else tried to influence her recollection, that is absolutely false and a mischaracterization of the facts*. Duration of *want*: <day.
- *Nationalists* <u>*want*</u> *to move towards Irish unity and see this process as a bridge in that direction*. Duration of *want*: ≥day.

4 Experiments and Results

We experiment with traditional SVM and neural networks. Our rationale behind SVM is to (a) incorporate deeper linguistic features than previous work, and (b) establish a solid baseline. We experiment with neural networks to evaluate the ability of word embeddings and recurrent neural networks to capture the context required to determine event durations. Regarding SVM, we use scikit-learn (Pedregosa et al., 2011). Regarding neural networks, we use Keras (Chollet et al., 2015) with TensorFlow backend (Abadi et al., 2015). All networks use GloVe embeddings with 300 dimensions (Pennington et al., 2014) and the Adam optimizer (Kingma and Ba, 2014). We use grid search and 5-fold cross-validation to tune hyperparameters (C and γ for SVM, and batch size, dropout rate, etc. for neural networks).

4.1 Support Vector Machine

Table 1 describes the full feature set. We use spaCy[1] to tokenize the input text and extract lemmas, part-of-speech tags, named entities, and dependencies. The features by Pan et al. (2011) and Gusev et al. (2011) capture primarily lexical information, relying on tendencies of particular words to denote events of certain durations. These tendencies are, however, subject to contextual influence. Duration is one component of the internal temporal structure of events, and as such it is an important factor for distinguishing between various aspectual categories (Vendler, 1957; Smith, 1991). It thus stands to reason that other features which capture aspectual distinctions may also correlate with event duration and be useful for classifying the duration of events in texts. In order to explore this intuition, we adapt features from a system designed to assign situation entity types to clauses (Friedrich et al., 2016). Diagnostic criteria for situation entity types include lexical aspect (stative vs. dynamic) of the main verb, genericity of the clause's subject, and whether the clause

[1] https://github.com/explosion/spaCy

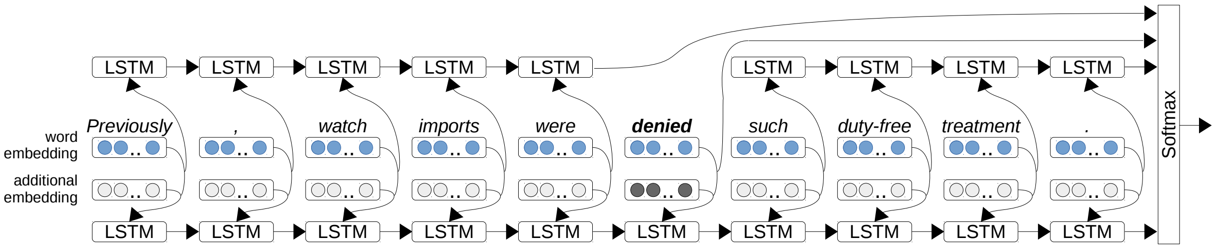

Figure 1: Neural network architecture to predict event durations. The output layer combines (a) the embedding of the verb at hand and (b) the output of three LSTMs: one for the whole sentence (bottom), one for the tokens before the event (top left), and one for the tokens after the event (top right).

is episodic, habitual, or static. It is primarily these criteria which features 26-50 aim to capture. For example, bare plural subjects with a simple present tense verb (e.g., *Bats eat mosquitos*) are a hallmark of GENERIC clauses. Although situation entity types do not directly map onto the duration labels (<day or ≥day), the criteria which contribute to determining them clearly influence aspectual interpretation, thus influencing understanding of the duration of events. Regarding Brown clusters, we use freely available clusters trained on news data by Turian et al. (2010) using the implementation by Liang (2005). We include one feature per cluster and set it to true if any word in the sentence belongs to the cluster.

4.2 Feed-Forward Neural Network

The first neural network we experiment with is a one-hidden-layer feed-forward neural network that takes as input the event embedding. The tuning process revealed that the size of the hidden layer is not important, thus we report results using a hidden layer with 5 neurons. Intuitively, this vanilla network evaluates whether pretrained word embeddings can predict the duration of events.

4.3 LSTM Ensemble

The LSTM ensemble is an improvement of the vanilla feed-forward neural network. It combines the event embedding with three LSTMs (Hochreiter and Schmidhuber, 1997) capturing different context around the event (Figure 1). The first LSTM (200 units, bottom in Figure 1) take as inputs the full sentence, and each token is represented by two embeddings: the word embedding (blue in Figure 1) and an additional embedding indicating whether the token is the event of interest or not (light and dark grey). The other two LSTMs (200 units each, top in Figure 1) take as input the sequence of tokens before and after the event at hand, respectively, and each token is represented by the corresponding word embedding. Word embeddings remain fixed, but the additional embeddings are initialized randomly and tuned during training along with all other network parameters.

		P	R	F1
Pan et al.	<day	.76	.57	.65
	≥ day	.70	.85	.77
	Avg.	.73	.72	.71
Pan et al. + Gusev et al.	<day	.73	.52	.61
	≥ day	.68	.84	.75
	Avg.	.70	.69	.68
Pan et al. + Gusev et al. + Situation Entities	<day	.82	.63	.71
	≥ day	.74	.89	.81
	Avg.	.78	.77	*.76
Feed-forward neural network	<day	.87	.63	.73
	≥day	.77	.93	.84
	Avg.	.81	.80	*.80
LSTM ensemble	<day	.97	.62	.76
	≥day	.78	.99	.87
	Avg.	.86	.83	*.82

Table 2: Results obtained using SVM and several feature combinations (top), and neural networks (bottom). We indicate statistical significance with respect to Pan et al. (2011) with *. Avg. stands for weighted average.

4.4 Results

Table 2 presents results obtained with the test set (WSJ data with 156 event instances). We used the same train and test splits as Pan et al. (2011) and Gusev et al. (2011), but reimplemented their systems and obtained better results than those reported by the authors. We believe this is due to the fact that spaCy (and the state-of-the-art in general) is more robust than older tools. Regarding SVM, the feature sets previously proposed obtain moderate results (F1: 0.71 and 0.68). These previous features clearly benefit from the new aspectual features (F1: 0.76), showing that the latter features capture contextual information useful to determine event durations. The feed-forward neural network outperforms the SVM (F1: 0.80) although it doesn't have access to the context surrounding the event at hand. This shows that embeddings alone are effective at predicting event durations. Finally, despite the relatively small dataset, the LSTM ensemble complements the pretrained verb embedding with distributional representations of the context around it (the full sentence, and the words before and after the event), yielding an 0.82 F1.

5 Error Analysis

In this section, we provide insights into why the additional aspectual features and neural networks are useful to predict event durations.

Aspectual features yield 7% improvement in overall F1 (0.71 vs. 0.76). Here are some examples that benefit from these features:

- *The company said 80% of its auction business is usually conducted in the second and fourth quarters*. The *adverbial degree* feature (feature 49) characterizes that *conducted* is a habitual event and made the SVM correctly classify this event into ≥day.
- *Nationalists want to move towards Irish unity and see this process as a bridge in that direction*. The subject of *want* is the bare plural *Nationalists* (feature 31), which in turn indicates that the event duration is ≥day.
- *Sotheby's Holdings Inc., the parent of the auction house Sotheby's, said its net loss for the seasonally slow third quarter narrowed from a year earlier on a leap in operating revenue*. The event *narrowed* belongs to the WordNet lexical filename *verb.change* and its object (*loss*) belongs to *noun.possession*. These semantic classes (features 41–43) made the classifier correctly predict ≥day. Another important lexical filename is *verb.possession*, all events belonging to this filename are annotated ≥day.

Neural Networks outperform any feature combination despite not having explicit access to any information beyond the sentence to which the event belongs and pretrained word embeddings. Word embeddings alone are surprisingly effective for this task (feed-forward neural network F1: 0.80), and benefit especially when the event at hand has not been seen in training. Similar to the WordNet lexical filenames, embeddings cluster together events with similar durations. The benefit of embeddings is, however, that they are pretrained on massive amounts of data and virtually account for any event (all the events annotated in the corpus we work with have a GloVe embedding). Here is an example of an unseen event in training that the embeddings predict correctly:

- *Revenue totaled $1.01 billion, a 43% increase from $704.4 million, reflecting the company's acquisition of Emery earlier this year.* The feed-forward neural network and embeddings learnt that mathematical expressions last less than a day (<day).

Although the difference in F1 is small (0.82 vs. 0.80), the LSTM ensemble successfully captures context required to predict event durations. Here are two examples that benefit:

- *The Portland, Ore., thrift said the restructuring should help it meet new capital standards from the Financial Institution Reform, Recovery and Enforcement Act.* The fact that *restructuring* appears nearby and has duration ≥day helps the LSTM ensemble predict that *meet* also has duration ≥day in this context, despite most meetings lasting less than a day. Also, the LSTM ensemble has access only to the nearby events but not to their duration.
- *In over-the-counter trading yesterday, Benjamin Franklin rose 25 cents to $4.25* (duration: <day). The LSTM ensemble is very successful when temporal cues surrounding the event at hand are present (e.g., *yesterday*).

6 Conclusions

In this paper, we classify events into those whose duration is shorter than a day (<day) or a day or longer (≥day). We have presented aspectual features that account for deeper linguistic information than previous work, and showed that they complement basic features used previously. We have also experimented with neural networks, and showed that (a) pretrained word embeddings successfully solve this task, and (b) an LSTM ensemble captures relevant context around the event despite that the corpus we work with is relatively small. We believe that determining the duration of events has the potential to help temporal reasoning in general. For example, somebody can participate in two events taking place at different locations only if they do not overlap temporally.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland. Association for Computational Linguistics.

François Chollet et al. 2015. Keras. `https://github.com/fchollet/keras`.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768, Berlin, Germany. Association for Computational Linguistics.

Andrey Gusev, Nathanael Chambers, Pranav Khaitan, Divye Khilnani, Steven Bethard, and Dan Jurafsky. 2011. Using query patterns to learn the duration of events. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, pages 145–154, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Percy Liang. 2005. *Semi-supervised learning for natural language*. Ph.D. thesis, Massachusetts Institute of Technology.

Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. 2015. Semeval-2015 task 5: Qa tempeval - evaluating temporal information understanding with question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 792–800, Denver, Colorado. Association for Computational Linguistics.

Feng Pan, Rutu Mulkar-Mehta, and Jerry R Hobbs. 2011. Annotating and learning event durations in text. *Computational Linguistics*, 37(4):727–752.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. 2003. The TIMEBANK corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656, Lancaster.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

James Pustejovsky, Mark Verhagen, Roser Saurí, Jessica Littman, Robert Gaizauskas, Graham Katz, Inderjeet Mani, Robert Knippen, and Andrea Setzer. 2006. Timebank 1.2. Linguistic Data Consortium.

Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. 2016. Temporal anchoring of events for the timebank corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2195–2204, Berlin, Germany. Association for Computational Linguistics.

Carlota S Smith. 1991. *The parameter of Aspect, vol. 43 of Studies in Linguistics and Philosophy*. Kluwer, Dordrecht.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Zeno Vendler. 1957. Verbs and times. *The philosophical review*, pages 143–160.

Jennifer Williams and Graham Katz. 2012. Extracting and modeling durations for habits and events from twitter. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 223–227, Jeju Island, Korea. Association for Computational Linguistics.