Timeline extraction using distant supervision and joint inference

Savelie Cornegruta
Department of Biomedical Engineering
King's College London, UK
savelie.cornegruta@kcl.ac.uk

Andreas Vlachos
Department of Computer Science
University of Sheffield, UK
a.vlachos@sheffield.ac.uk

Abstract

In timeline extraction the goal is to order all the events in which a target entity is involved in a timeline. Due to the lack of explicitly annotated data, previous work is primarily rule-based and uses pre-trained temporal linking systems. In this work, we propose a distantly supervised approach by heuristically aligning timelines with documents. The noisy training data created allows us to learn models that anchor events to temporal expressions and entities; during testing, the predictions of these models are combined to produce the timeline. Furthermore, we show how to improve performance using joint inference. In experiments in the SemEval-2015 TimeLine task we show that our distantly supervised approach matches the state-of-the-art performance while joint inference further improves on it by 3.2 F-score points.

1 Introduction

Temporal information extraction focuses on extracting relations and events along with the time when they were true or happened. In this work we focus on timeline extraction, following the recent SemEval TimeLine shared task (Minard et al., 2015). The aim of the task is to extract timelines from multiple documents consisting of events in which a given target entity is the main participant. An example timeline for the entity *Steve Jobs* extracted from 4 documents is given in Fig.1.

The development data provided by the TimeLine shared task does not contain annotations for the various intermediate processing stages needed, only a set of documents with annotated event mentions (input) and the timelines extracted for a few target entities (output). No training data was provided, thus participating systems used rules combined with temporal linking systems trained on related tasks in order to anchor events to temporal expressions and entities to construct the timelines.

We propose a new approach to timeline extraction that uses the development data provided as distant supervision to generate noisy training data (Craven and Kumlien, 1999; Mintz et al., 2009). More specifically, we heuristically align the target entity and the timestamps from the timelines with automatically recognized entities and temporal expressions in the documents. This noisy labeled data set allows us to learn models for the subtasks of anchoring events to temporal expressions and to entities, without requiring training models on additional data. Also, we improve the performance using joint inference for both anchoring subtasks. In our experiments, we show that our distantly supervised approach matches the state-of-the-art performance while joint inference further improves on it by 3.2 F-score points. Our code is publicly available at `http://github.com/savac/timeline`.

2 Timeline extraction

The task of timeline extraction given a target entity and a set of documents can be decomposed as follows. The initial stages are event mention extraction, target entity recognition, and temporal expression identification and resolution. The next stages are anchoring event mentions to target entities and temporal expressions. The final stages are event corefer-

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1936–1942,
Austin, Texas, November 1-5, 2016.

Documents:

DocId: 16844, DCT: 2010-06-08, Sentence: 2,3,4,5: *Yesterday*[2010-06-07] *, at this year 's Apple Worldwide Developers Conference (WWDC) , company CEO Steve Jobs*[Steve Jobs] *unveiled iPhone 4*[iPhone 4] *, along with the new iOS 4 operating system for Apple mobile devices . The announcement was long-awaited but not a very big surprise . In April , the technology blog Gizmodo obtained a prototype of the new phone*[iPhone 4] *and published details of it*[iPhone 4] *online . While introducing iPhone 4*[iPhone 4] *, at the annual conference , Jobs*[Steve Jobs] *started by hinting at the incident , saying , " Stop me if you 've already seen this .*

DocId: 17036, DCT: 2010-07-17, Sentence: 6,15: *Rather than recall the devices or offer a hardware fix , Jobs*[Steve Jobs] *said yesterday*[2010-07-16] *that Apple will offer a free case to anyone who has purchased an iPhone 4*[iPhone 4]*. [...] However Jobs*[Steve Jobs] *admitted that the percentage of calls dropped on the iPhone 4*[iPhone 4] *was slightly greater than the percentage of calls dropped on the 3GS*[iPhone 3GS]*.*

DocId: 16900, DCT: 2010-06-16, Sentence: 6: *The newest iPhone*[iPhone 4] *, iPhone 4*[iPhone 4] *was introduced by Apple CEO*[Steve Jobs] *Steve Jobs*[Steve Jobs] *at the company 's 2010 Worldwide Developer 's Conference less than two weeks ago*[2010-06] *.*

DocId 16983, 2010-10-23, Sentence 10: *In his*[Steve Jobs] *keynote address Wednesday*[2010-10-20] *, Jobs*[Steve Jobs] *announced the release of Apple 's iLife '11 software suite , which includes the iPhoto , iMovie , and GarageBand programs .*

Timeline: *Steve Jobs*

1	2010-06-07	16844-2-unveiled	
1	2010-06-07	16844-5-introducing	16900-11-introduced
1	2010-06-07	16844-5-hinting	
1	2010-06-07	16844-5-saying	
2	2010-07-16	17036-6-said	
2	2010-07-16	17036-15-admitted	
3	2010-10-20	16983-10-address	
3	2010-10-20	16983-10-announced	

Figure 1: Example timeline for target entity *Steve Jobs*. The input to the system is the documents annotated with event mentions annotations and their Document Creation Time (DCT). The event mentions appearing in the timeline are identified by their document id-sentence index. The annotations for the target entities and temporal expression mentions need to be done by the system.

ence resolution and ordering of the events in a timeline, which rely largely on their anchoring to temporal expressions. The TimeLine shared task had two tracks, A and B, the only difference being that in Track B the event mentions are provided in the input. We consider this track in this paper and focus on learning the anchoring of events to temporal expressions and entities.

The development data provided in the context of the shared task consisted of documents related to *Apple* and gold timelines for six target entities. Evaluation was performed by extracting timelines from three document sets, each related to *Airbus*, *GM* and *Stock market* respectively. We used the official evaluation which is based on the metric introduced by UzZaman and Allen (2011) which assesses a predicted timeline versus the gold standard one using precision, recall and F-score over binary temporal relations between the events.

3 Distant supervision

In order to generate training data for anchoring event mentions to target entities and temporal expressions via distant supervision, we first need to identify them. For entity recognition we use approximate string matching combined with the Stanford Coreference Resolution System (Lee et al., 2013). For temporal expression identification and resolution to absolute timestamps we use the UWTime temporal parser (Lee et al., 2014).

Next we generate labeled instances as follows. For anchoring events to entities, we consider for each event mention the correct entity mention to be the nearest mention of the target entity in the same sentence, and all others to be incorrect. Similarly, for anchoring events to timestamps, we consider for each event mention the correct temporal expression to be the nearest temporal expression that exactly matches the timestamp according to the timeline (but not necessarily in the same sentence), and all others to be incorrect. The datasets generated will be noisy since correct anchors may be entity mentions and temporal expressions that are not the nearest ones. Further noise is expected due to errors in the entity recognition and temporal expression identification and resolution stages.

Features	type
Measure distance in tokens between event and target entity mentions	local
Syntactic dependencies between event and target entity mentions (extracted from training corpus)	local
Check if subsequent events have the same stem and are attributed to the same target entity	global
Check if subsequent events are in the same sentence and are attributed to the same target entity	global
Check if subsequent events are both communication events and are attributed to the same target entity	global

Table 1: Features to encode dependencies between events and target entities

4 Event anchoring

After generating training data for anchoring event mentions to target entities and to temporal expressions with distant supervision, we now proceed to developing linear models for each of these tasks.

4.1 Classification

Using distant supervision we obtained examples of correct and incorrect anchoring of event mentions to entities and temporal expressions. Thus we learn for each of the two tasks a binary linear classifier of the form:

$$score(x, y, \mathbf{w}) = \mathbf{w} \cdot \phi(x, y) \qquad (1)$$

where x is an event mention, y is the anchor (either the target entity or the temporal expression) and $\mathbf{w}$ are the parameters to be learned. The features extracted by ϕ represent various distance measures and syntactic dependencies between the event mention and the anchor obtained using Stanford CoreNLP (Manning et al., 2014). The temporal expression anchoring model also uses a few feature templates that depend on the timestamp of the temporal expression. The full list of features extracted by ϕ are denoted as local in Tables 1 and 2.

4.2 Alignment

The classification approach described is limited to anchoring each event mention to an entity or a temporal expression in isolation. However it would be preferable to infer the decisions for each task jointly at the document level and take into account the dependencies in anchoring different events, e.g. that consecutive events in text are likely to be anchored

Features	type
Measure distance in sentences between event mention and temporal expression	local
Measure distance in tokens between event mention and temporal expression	local
Syntactic dependencies between event mention and temporal expression (extracted from training corpus)	local
Check if temporal expression is before of after the event mention	local
Check if timestamp is in the future wrt the DCT	local
Check if timestamp is undefined (i.e. XX-XX-XXXX)	local
Check if timestamp is incomplete	local
Check if subsequent events and are linked to the same temporal expression	global
Check if subsequent events have the same stem and are linked to the same temporal expression	global
Check if subsequent events are in the same sentence and are linked to the same temporal expression	global
Check if subsequent events are communication events and are linked to the same temporal expression	global

Table 2: Features to encode dependencies between events and temporal expressions

to the same entity, as shown in Figure 2, or to the same temporal expression. Capturing such dependencies can be crucial when the correct anchor is not explicitly signalled in the text but can be inferred considering other relations and/or their ordering in text (Derczynski, 2013).

Defining our joint model formally, let **x** be a vector containing all event mentions in a document and **y** be the vector of all anchors (target entity mentions or temporal expressions) in the same document. The order of the events in **x** is as they appear in the document. Let **z** be a vector of the same length as **x** that defines the alignment between **x** and **y** by containing pointers to elements in **y**, thus allowing for multiple events to share the same anchor. The scoring function is defined as

$$score(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}) = \mathbf{w} \cdot \Phi(\mathbf{x}, \mathbf{y}, \mathbf{z}) \qquad (2)$$

where the global feature function Φ, in addition to the features returned by the local scoring function (Eq. 1), also returns features taking into account anchoring predictions across the document. Apart from features encoding subsequences of anchoring

Documents:

DocId: 17036, DCT: 2010-07-17, Sentence: 9,10,11:

Jobs[Steve Jobs] *also* *said* *that those who had already purchased a bumper will receive a full refund for the accessory. For consumers still dissatisfied with* *iPhone 4*[iPhone 4], *Jobs*[Steve Jobs] *said* *that the phones can be returned for a refund as well.* *Jobs*[Steve Jobs] *acknowledged* *that "a very small percentage of users" were experiencing antenna issues, but* *dismissed* *the existence of an "Antennagate,"* *saying* *that similar problems plague all cellular phones and that the* *iPhone*[iPhone 4] *issue* *"has been blown so out of proportion that it is incredible."*

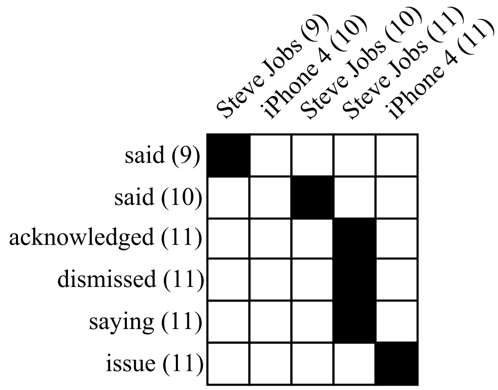

Figure 2: The correct alignment of events and target entity mentions is shown with the numbers in brackets denoting the index of the sentence in which the mention is found. The consecutive events acknowledged, dismissed and saying are anchored to entity Steve Jobs that was only mentioned once in the beginning of the sentence.

predictions, it also makes possible to make them dependent on the events, e.g. a binary indicator encoding whether two consecutive events with the same stem share the same anchor or not. The full list of local and global features extracted by Φ are presented in Tables 1 and 2. Predicting with the scoring function in Eq.2 amounts to finding the anchoring sequence vector $\mathbf{z}$ that maximizes it. To be able to perform exact inference efficiently, we impose a first order Markov assumption and use the Viterbi algorithm (Viterbi, 1967). Similar approaches have been successful in word alignment for machine translation (Blunsom and Cohn, 2006).

4.3 Post-processing

During testing, we need to construct the timeline for each target entity using the events that were predicted to be anchored to it and the timestamps of the temporal expressions each event was anchored to. Thus, we need to perform two additional tasks, event coreference and ordering. For the former we define a simple heuristic by which if two mentions have the same stems and timestamps then they refer to the same event. The only exception is that if two mentions represent communication events (*said*, *announced* etc.), then they are resolved to different events when in the same document. We finally order the events according to their timestamp.

5 Results

We evaluate our system using the setup provided by the TimeLine task ensuring that the training and validation are performed only using the development data i.e. the *Apple* collection. All linear models were trained with the perceptron update rule (Pedregosa et al., 2011). We tuned the number of perceptron iterations by performing cross-validation using the development data by holding out the timeline for one target entity and training on the timelines for the remaining ones.

In Table 3 we compare the binary classification model (Our_System_Binary) against the alignment model (Our_System_Alignment) and show that the latter outperforms the former by a margin of 3.2 points in F-score, achieving a micro F_1-score of 28.58 across the three test corpora, thus confirming the benefits of joint inference. The only corpus in which joint inference did not help was *Stock* which has on average shorter event chains per document (Minard et al., 2015) and thus renders joint anchoring less likely to be useful.

We now compare our approach to the two participants in the TimeLine shared task with two runs each. The best-performing GPLSIUA team (Navarro and Saquete, 2015) used the TIPSem tool developed by Llorens et al. (2010) for temporal relation processing which extracts events and temporal expressions and uses a Conditional Random Field model to anchor them against each other. However, TIPSem only considers anchoring of events to temporal expressions that are in the same sentence. GPLSIUA also used the semantic role labeler from SENNA (Collobert et al., 2011) and OpenNER and anchored entities to events using a rule-based approach. The HeidelToul team (Moulahi et al., 2015) used HeidelTime (Strötgen et al., 2013) to identify and resolve temporal expressions and de-

System	Airbus F$_1$	GM F$_1$	Stock F$_1$	Total P	Total R	Total F$_1$
GPLSIUA_1	22.35	19.28	33.59	21.73	**30.46**	25.36
GPLSIUA_2	20.47	16.17	29.90	20.08	26.00	22.66
HeidelToul_1	19.62	7.25	20.37	20.11	14.76	17.03
HeidelToul_2	16.50	10.82	25.89	13.58	28.23	18.34
Our_System_Binary	17.99	20.97	**34.95**	25.97	24.79	25.37
Our_System_Alignment	**25.65**	**26.64**	32.35	**29.05**	28.12	**28.58**

Table 3: Results for our system and other participants in the SemEval 2015 Task 4: TimeLine.

veloped a target entity mention identification tool similar to ours using Stanford CoreNLP (Manning et al., 2014). However, they rely on a rule-based approach for event anchoring. Our binary model matches the performance of the best system, and our alignment model exceeds it by 3.2 F_1-score points across, even though we do not use any off-the-shelf components developed for temporal relation extraction. Instead we rely on training data generated with distant supervision, and UWTime for temporal expression identification and resolution, for which the participants also used similar components.

6 Related work

In recent work, Laparra et al. (2015) also considered anchoring at the document-level in the context of the Track A of the TimeLine shared task, however they developed a rule-based approach. The structure features used in our joint inference approach encode similar intuitions, but we are learning model weights using distant supervision so that we can combine them more flexibly. And even though the noise in the trainng data generated with distant supervision is a concern, manual annotation of temporal relations is known to have low inter-annotator agreement rates[1] and thus also likely to be noisy.

Prior to the TimeLine shared task, TempEval (Verhagen et al., 2007) was the original task that focused on categorising the relations between events, temporal expressions and Document Creation Time using the the TimeML annotation language. The task classified only the relations between mentions in the same or consecutive sentences. The two following tasks, TempEval-2 (Verhagen et al., 2010) and TempEval-3 (UzZaman et al., 2013), added tasks for event and temporal expression identification as well as an end-to-end temporal relation processing task that was performed on raw text.

Beyond TempEval, McClosky and Manning (2012) used distant supervision in order to learn how to extract the temporal bounds for events in the context of the TAC temporal knowledge base population task (Ji et al., 2011). However they focus on learning real-world event ordering constraints (e.g. people go to school before university) instead of how events are reported in text.

7 Conclusions

In this paper we proposed a timeline extraction approach in which we generate noisy training data for anchoring events to entities and temporal expressions using distant supervision. By learning a binary classifier we match the state-of-the-art F_1-score for the Track B of the TimeLine shared task. We further improve this result by 3.2 F_1-score points using joint inference.

Acknowledgments

Part of this work was conducted while both authors were at University College London. The authors would like to thank Leon Derczynski for his feedback on an earlier version. Andreas Vlachos is supported by the EU H2020 SUMMA project (grant agreement number 688139).

References

Phil Blunsom and Trevor Cohn. 2006. Discriminative word alignment with conditional random fields. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 65–72. Association for Computational Linguistics.

[1] `http://www.timeml.org/timebank/documentation-1.2.html`

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 77–86. AAAI Press.

Leon Derczynski. 2013. *Determining the Types of Temporal Relations in Discourse*. Ph.D. thesis, University of Sheffield.

Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2011. Overview of the tac 2011 knowledge base population track. In *Proceedings of Text Analysis Conference (TAC)*.

Egoitz Laparra, Itziar Aldabe, and German Rigau. 2015. Document level time-anchoring for timeline extraction. In *ACL*.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Comput. Linguist.*, 39(4):885–916, December.

Kenton Lee, Yoav Artzi, Jesse Dodge, and Luke Zettlemoyer. 2014. Context-dependent semantic parsing for time expressions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447, Baltimore, Maryland, June. Association for Computational Linguistics.

Hector Llorens, Estela Saquete, and Borja Navarro. 2010. Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

David McClosky and Christopher D Manning. 2012. Learning constraints for consistent timeline extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 873–882. Association for Computational Linguistics.

Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, and Ruben Urizar. 2015. Semeval-2015 task 4: Timeline: Cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 778–786, Denver, Colorado, June. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bilel Moulahi, Jannik Strötgen, Michael Gertz, and Lynda Tamine. 2015. Heideltoul: A baseline approach for cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 825–829, Denver, Colorado, June. Association for Computational Linguistics.

Borja Navarro and Estela Saquete. 2015. Gplsiua: Combining temporal information and topic modeling for cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 820–824, Denver, Colorado, June. Association for Computational Linguistics.

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.

Jannik Strötgen, Julian Zell, and Michael Gertz. 2013. Heideltime: Tuning english and developing spanish resources for tempeval-3. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 15–19, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Naushad UzZaman and James F. Allen. 2011. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 351–356, Stroudsburg, PA, USA. Association for Computational Linguistics.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80. Association for Computational Linguistics.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62. Association for Computational Linguistics.

Andrew J. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, IT-13(2):260–269, April.