Extracting Complex Information from Natural Language Text: A Survey

Emna Mechket, Fabian Suchanek

Extracting Complex Information from Natural Language Text: A Survey

Mechket Emna Mahouachi[a], Fabian Suchanek[b]

[a] ENSTA Paris, Institut polytechnique de Paris, France
[b] Telecom Paris, Institut polytechnique de Paris, France

Abstract
Information Extraction is the art of extracting structured information from natural language text, and it has come a long way in recent years. Many systems focus on binary relationships between two entities – a subject and an object. However, most natural language text contains complex information such as beliefs, causality, anteriority, or relationships that span several sentences. In this paper, we survey existing approaches at this frontier, and outline promising directions of future work.

Keywords
Information Extraction, Semanting parsing, Complex Information

1. Motivation

Information extraction (IE) is the process of extracting machine-readable, structured information from natural language text. For example, given the sentence "Angelina Jolie stars in the superhero film The Eternals", an IE system can extract the fact ⟨*Angelina Jolie, stars, The Eternals*⟩. IE finds numerous applications, be it in search engines, science, or the digital humanities [1]. In journalism, for example, IE techniques have been used for fact checking [2], to examine the "Panama Papers" [3], or to extract semantic information from web pages (as in Reuter's Calais service).

Most of these IE systems extract *triples*, i.e., facts that consist of a subject, a predicate, and an object. In our example ⟨*Angelina Jolie, stars, The Eternals*⟩, the subject is *Angelina Jolie*, the predicate is *stars*, and the object is *The Eternals*. However, much of the information that we care about is not of this form. Consider for example the following sentence (taken from the Wikipedia article about Angelina Jolie): "Jolie applied for adoption as a single parent, because Vietnam's adoption regulations do not allow unmarried couples to co-adopt". This sentence does not talk about a simple triple. Instead, it contains a negation, a modifier ("as a single parent"), and a causal relationship. A cursory reading of any Wikipedia article, blog, journalistic piece of text, or even just the present paper suggests that the majority of sentences is not concerned with simple triples, but with more complex information.

The question thus arises to what degree current IE systems can deal with such information. In this survey article, we focus on 5 systems that have particular provisions for dealing with more than triples: FRED [4], K-Parser [5], ClausIE [6], MinIE [7], and OpenIE [8]. We systematically analyze their ability to extract different types of complex information. We study the dimensions that have been identified as particularly challenging [9]: beliefs, negation, causality, anteriority, *n*-ary relations, cross-sentence references, and contrast. Finally, we outline the open challenges in the area. Our survey is structured as follows: In Section 2, we discuss the different dimensions of knowledge that we are interested in, and our test set of sentences. Section 3 presents the systems that we study, and their results on our test sentences. Section 4 summarizes our findings, before Section 5 concludes.

2. Dimensions of complex knowledge

Numerous surveys discuss information extraction systems (e.g., [10, 11, 12, 13]). While [13] also studies inter-proposition relationships, to the best of our knowledge, no survey has yet systematically compared the performance of IE systems for complex knowledge that goes beyond triples. In this survey, we focus on the following dimensions (loosely based on [9]):

Belief. We are interested in sentences where a subject expresses a belief in a hypothetical statement, as in "Researchers believe that the Corona virus will disappear". An IE system shall extract that researchers have a certain belief, but the system shall not extract that the Corona virus will de facto disappear (since this is just a hypothetical scenario).

Negations are sentences that contain a negation particle, such as "Brad Pitt and Angelina Jolie are no longer

Proceedings of the CIKM 2020 Workshops, October 19-20, 2020, Galway, Ireland Editors of the proceedings (editors): Stefan Conrad, Ilaria Tiddi
mechket-emna.mahouachi@ensta-paris.fr (M.E. Mahouachi); suchanek@telecom-paris.fr (F. Suchanek)

together". The IE system should explicitly pinpoint the negative polarity of the sentence.

Causality expresses a causal relationship between two events, as in "The SpaceX rocket launch has been postponed because of the bad weather". We want the IE system to extract the two events, and to link them by a causal relation.

Anteriority expresses that one event happened before another one, as in "Once the votes are closed, the new president will be known", or in "Monica took her driver's license before buying a car". Again, we want the IE system to extract two events and a link between them.

Contrast is expressed by conjunctions such as "although", and indicates that one event happened despite another one – as in "Even though the weather was bad, they enjoyed the concert". This constellation carries a slightly different meaning from the sentence that links the two propositions by a simple "and", and we want the system to mirror this.

N-ary relations are relations with more than two participants, as in "Emma bought a book from FNAC for $12 as a mother's day gift". Here, the goal is to extract the agent Emma, the patient (the book), a purpose, the price, and the provenance of the book – in the spirit of frames [14].

Anaphoras are words (such as "she") that refer to an entity that has already been mentioned, as in "Walter Elias Disney was passionate about drawing since he was young". Here, we want the system to link "he" to Disney.

Inter-sentence relations are relations between events that are expressed in different sentences. For example, we can say "Covid-19 spread around the world. For this reason, many countries went into lockdown". Here, we expect the system to extract a causal relationship between the two sentences. The same principle can be applied to causality, anteriority, *n*-ary relations, contrast, and anaphoras.

Note that we do not expect the output of the system to be of a certain shape; we just want the aspect of knowledge to be mirrored in some way in the output of the system. To systematically analyse the different IE systems, we prepared 3-5 test sentences for each of our dimensions of complex knowledge. Our sentences vary the order, the topic, and the syntactic devices. For example, for causality, we have sentences that use "because" followed by a noun phrase, or by a subordinate clause. Table 1 shows some example sentences. Our full set of test sentences is available at our project Webpage https://github.com/michka2/Complex-IE.

3. IE Systems

We focus on IE systems that aspire to extract more than simple triples. We found the following state-of-the-art systems: ClausIE, MinIE, and OpenIE, K-Parser, and FRED. These systems fall into two categories: *Semantic parsers* extract a structure from the input sentence that mirrors semantic relationships such as "agent of". Representatives of these systems are FRED and K-Parser. Then there are *Information Extraction Systems* in the proper sense of the word, which extract facts in a quasi-logical representation. We study the systems ClausIE, MinIE, and OpenIE.

Another distinction is between *open* and *closed* IE systems. Closed IE systems aim to extract facts whose components are from a predefined catalog of entities and relations. For example, from "Jolie will appear in The Eternals", they aim to extract ⟨*Angelina_Jolie, stars, The_Eternals_(2020_movie)*⟩ – where "Jolie" has been mapped to the entity *Angelina_Jolie*, the "will appear" is mapped to the predefined relation *stars*, and the movie is identified unambiguously. This is what the FRED system does. Open systems, in contrast, use spans of the input sentence as subject, predicate, and object of the triple – as in ⟨*Jolie, will appear in, The Eternals*⟩. This is what ClausIE, MinIE, K-Parser, and the eponymic OpenIE systems do. There are near-philosophical debates about which of the two paradigms is better suited for IE. Often, the outputs of a closed IE system are better for reasoning and querying (because different surface forms of the same entity have been canonicalized), but open IE systems can extract information from a wider variety of sentences (because they are not limited to the predefined relations). Here, we just note that the choice of *open* versus *closed* is a dimension that is orthogonal to the issues that we study in this survey.

We shall now present each system in detail, comment on its usability, and investigate how it deals with our test sentences.

3.1. ClausIE

ClausIE (Clause-Based Open Information Extraction [6]) is an open information extraction system. Each extracted fact contains a subject, a relation, and optionally one or several arguments. The system uses dependency parsing to identify the *clauses* in the sentence, i.e., the parts that express a coherent piece of information. Each clause is then transformed into a fact. The system can be found online[1], and is easy to run. Let us now study how ClausIE performs on our test set.

Beliefs. ClausIE deals with this kind of sentences by generating two independent propositions. For exam-

[1] https://github.com/IsaacChanghau/ClausIE

Causality	The SpaceX rocket launch has been postponed because of bad weather.
Anteriority	Once the votes are closed, the new president will be known.
Negation	It is not necessary to sanitize your groceries.
n-ary	Angelina Jolie has been named Hollywood's best paid actress.
Anaphoras	Walter Disney was passionate about drawing since he was young.
Contrast	Even though the weather was bad, they enjoyed the concert.
Multi-sentence	Lady Diana died in a car crash. Since that night, many people believe that her death was ordered by the Royal Family because of her relationship with Dodi Fayed.

Table 1
Some of our test sentences.

ple, from "Aristotle thought that the Sun moves around the Earth", ClausIE will extract:

("Aristotle", "thought","that the Sun moves around the Earth")
("the Sun", "moves", "around the Earth")

These are indeed propositions in ClausIE's sense, but the hypothetical nature of the second one is not mirrored in the output.

Negation. ClausIE extracts the negative verb as the relation between the two arguments, as shown here:

("Ross and Rachel", "were not", "on a break")

Thus, the identification of the negation is left to the user.

Causality. ClausIE does not create a particular link between two causally related propositions. We show here the result of "Trump wants to ban Twitter because of the criticism he is receiving":

("Trump", "wants","to ban Twitter because of the criticism he is receiving")
("he", "is receiving", "criticism")

The same goes for **anteriority** and **contrast**.

N-ary relations are a setting that ClausIE handles explicitly. It extracts one proposition for each contributing entity:

("Emma", "bought","a pride and prejudice book from fnac for $12")
("Emma", "bought","a pride and prejudice book as a mother's day gift")
("a mother", "has", "day gift")

ClausIE misinterpreted the "mother's day gift" as if the mother owned a "day gift". However, the first line of the output correctly shows that there is one main phrase, where the parts with "from FNAC" and "As mother's day gift" are optional (shown by the question mark).

Anaphoras are left as is by ClausIE, as shown in the parsing of "Scientists ignored Einstein at first, thinking he lost his mind":

("Scientists", "ignored", "E. at first thinking he lost his mind")
("Scientists", "ignored", "Einstein at first)
("Scientists", "be thinking", "he lost his mind")
("he", "lost", "his mind")
("his", "has", "mind")

Inter-sentence relations. ClausIE can deal with only a single sentence at a time. Thus, it is unable to see connections between two sentences, be it causality, anteriority, or anything else.

Summary. ClausIE identifies clauses, i.e., pieces of a sentence that express a coherent piece of information. Its results project away the aspects of anteriority, causality, contrast, and belief. Anaphoras are not resolved. However, ClausIE can effectively deal with *n*-ary relations.

3.2. MinIE

MinIE (Minimizing Facts in Open Information Extraction [7]) is an open information extraction system based on ClausIE. It augments the output by information on polarity, modality, attribution, and quantities with semantic annotations. In return it removes parts that are considered overly specific. The system is available online[2], and is easy to run.

We used the "safe mode" of MinIE for our experiments, because it omits only very few pieces of information. On our test sentences, MinIE performs as follows:

Belief. MinIE deals particularly well with this kind of sentence. It detects verbs, adverbs and adjectives that express possibility and certainty, and it annotates the triples accordingly. For "Aristotle thought that the Sun moves around the Earth", we obtain

Triple: "Sun", "moves around", "Earth"
Factuality: (+, CT), Attribution: (Aristotle, (+, PS))

This extraction means that the sentence is positive ("+") and that it is a certainty ("CT"). The triple itself is also positive, attributed to Aristotle, and a possibility (as indicated by the nested "PS"). If we change the verb

[2]https://github.com/uma-pi1/minie

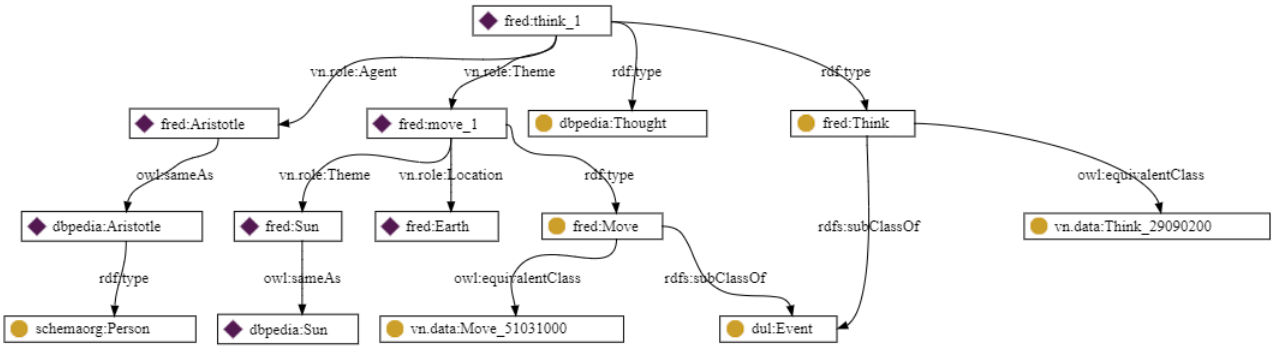

Figure 1: FRED output for a belief sentence.

from "thought" to "confirms", then the nested factuality changes to "CT" (for "certainty").
Negation is also a challenge where MinIE shines: It extracts the verb without negation, and changes the polarity from "+" to "-".
Causality is not explicitly treated in MinIE. For example, "The SpaceX rocket launch has been postponed because of the bad weather" yields:

Triple: "SpaceX rocket launch",
"has been postponed because of",
"bad weather"
Factuality: (+, CT), Attribution: none

Anteriority, likewise, is not mirrored explicitly in MinIE. When we test the sentence "Once the votes are closed, the new president will be known", MinIE does not detect the link between the two facts (the factualities are all (+,CT), with no attribution):

Triple: "Votes", "are closed", "Once"
Triple: "Votes", "are closed"
Triple: "new president", "be known"

The same goes for **contrast**.
N-ary relations are handled well by MinIE, since it is based on ClausIE. For our example sentence with the book from FNAC, we obtain:

Triple: "Emma", "bought a book from", "FNAC"
Triple: "Emma","bought a book from FNAC for", "QUANT_0_1"
Triple: "Emma","bought a book from FNAC as", "mother's day gift"
Triple: "mother", "has ", "day gift"

Anaphoras are left unlinked by MinIE, as by ClausIE.
Inter-sentence relations. Much like ClausIE, MinIE can deal with only a single sentence at a time.
Summary. MinIE is based on ClausIE, and thus shines natively on *n*-ary relations. Furthermore, it explicitly extracts polarity and attribution, two of our desiderata. Anaphoras, in contrast, are left unresolved. The causal and other links between two parts of a sentence are also not visible in MinIE's output.

3.3. OpenIE

The paradigm of Open Information extraction has been pushed forward by a series of systems from the University of Washington:

Textrunner [15], ReVerb [16], OLLIE [8], and the latest system, OpenIE 5 [8]. We focus on OpenIE 5, which is available online[3]. The tool is not easy to run: We have to download the build tool SBT[4], a Language Model (from a Google drive), and the actual OpenIE code from the git repository. These items have to be in specific folders, and the actual system runs as a Scala program by help of the Java runtime environment. Let us now see how OpenIE 5 works on our sample sentences:
Belief. OpenIE 5 can correctly identify nested sentences. For our example sentence "Aristotle thought that the Sun moves around the Earth", it extracts:

Context(Aristotle thought; List[0,17)):
(earth; moved; around the sun)
(Aristotle; thought; earth moved around the sun)

Here, OpenIE 5 creates a *context*, fills it with the hypothesis, and attributes it to Aristotle. This retains that the hypothesis is not asserted. The expression "List [0,17)" represents the location of the context in the input sentence.
Negation. OpenIE 5 deals with negative sentences by extracting the verb in its negative form as the relation part:

Angeline Jolie; is not; from Tunisia

Causality is represented in OpenIE 5 by an additional argument:

[3]https://github.com/dair-iitd/OpenIE-standalone
[4]https://www.scala-sbt.org/

Covid-19; cannot be compared; to previous pandemics; because medical research has advanced a lot.
medical research; has advanced; a lot.

The analysis thus boils down to a chunking of the input sentence.

Anteriority is represented by a special time-indicator in Open IE 5. For "Monica took her driver's license before buying a car", we obtain:

Monica; took; her driver's licence;T: before buying a car

N-ary relations are dealt with by creating multiple arguments for the output fact. Our example sentence with the book yields:

Emma; bought; a book; from FNAC; for $12 as a gift

However, with complex and long sentences, the system can get confused. It either omits relevant information or generates wrong tuples. For example, for "Michelle Obama is not only known as Obama's wife, but also as an advocate for poverty awareness", we obtain only:

Michelle Obama; is not known; as Obama's wife

Contrast just yields two independent triples in OpenIE 5. For "Despite the fact that the bus was late, we arrived on time", we obtain:

we; arrived; T: on time
the bus; was; late

Anaphoras are not handled specifically by Open IE 5. They just remain as unlinked pronouns in the triples.

Inter-sentence relations do not receive special treatment in OpenIE 5. The system just proceeds sentence by sentence.

Summary. OpenIE 5 shines with temporal expressions and belief sentences, for which it outputs special types of triples. For anaphoras, causality, contrast, and *n*-ary relations, Open IE 5 resorts basically to a chunking of the input sentence.

3.4. FRED

FRED [4] is a semantic parsing system that is particularly tailored to the Semantic Web. In this spirit, it produces a machine-readable RDF/OWL representation of the sentence, categorizes each entity into a set of predefined classes, disambiguates the mentioned entities to existing entities in the DBpedia knowledge base, and uses standard Semantic Web vocabulary wherever possible. For this purpose, the system combines a wide array of NLP and Semantic Web tools.

FRED is available online as a demonstration[5]. It can also be used programmatically through an API (with a key from the developers).

Belief. In FRED, a belief sentence gives rise to one main event (the act of believing), which has as object another event, the belief itself. Hence, "Aristotle thought that the Earth moved around the Sun" yields the parse shown in Figure 1. Here, "Think" is the main event, which has Aristotle as an Agent and "Move" as a theme. This theme is again an event, which has a theme and a location. All predicates are standard relations from the RDF, OWL, and VerbNet ("VN") vocabulary. Furthermore, all entities have been mapped to unambiguous DBpedia entities.

Negation is expressed by an additional relation "boxing: hasTruthValue" that leads to a node labeled with "boxing:False". (All parsing trees are available on our Web page.)

Causality between two events in a single sentence is expressed, quite naturally, by a relation "fred:because" between the main event and the event that is the reason for the main event. The same technique is used for **anteriority** and **contrast**.

N-ary relations are handled very well by FRED: Each participant of the event is linked by a dedicated relation to the main verb.

Anaphoras are not linked to their referent by FRED. The system creates a node for the pronoun, and labels it with "male" or "female", but does not establish the link to the referent.

Inter-sentence relations. FRED does not detect the link between two events across two sentences. It deals with each event separately. FRED is also not able to perform anaphora resolution across two sentences.

Summary. FRED can detect negation, and represent beliefs, causality, anteriority, and contrasting sentences. It also deals well with n-ary relations. However, it does not resolve anaphoras, and it does not detect relationships across sentences.

3.5. Knowledge Parser

K-parser [5] is a semantic parser, which transforms a natural language text into a machine-readable representation. It generates a graph similar to the one we have already seen for FRED. K-parser identifies event-event relations like causality or anteriority, event-entity relations such as semantic roles, and instance-of relations between entities and classes. The K-parser demonstration is no longer available online, but the code can be downloaded[6], and can be run by carefully following the instructions. Let us now study the performance of this system on our set of test sentences:

[5] http://wit.istc.cnr.it/stlab-tools/fred/
[6] https://github.com/arpit7123/K-Parser-JAR

Belief. For belief sentences, K-parser creates one event for each proposition – without detecting the link between them. For the sentence "Aristotle thought that the Earth moved around the Sun", K-parser yields:

has(Thought, Agent, Aristotle)
has(Aristotle, semantic_role, thinker)
has(moved, agent, Earth)
has(Earth, semantic_role, mover)
has(think, is_subclass_of, cognition)
has(move, is_subclass_of, motion)

Negation is specifically identified by K-Parser. It links the negated verb and the negation particle by the relation "negative". For example, for "Joe Tribuani doesn't speak French", we obtain:

has(speak, negative, not)
has(speak, is_subclass_of, communication)
has(not, is_subclass_of, all)
has(speak, is_subclass_of, communication)
has(does, instance_of, do)
has(Joe_Tribuani, is_subclass_of, person)
has(Joe_Tribuani-1, semantic_role, talker)
has(French, semantic_role, language)
has(do, is_subclass_of, social)
has(Speak, agent, Joe_Tribuani)
has(speak, recipient, French)

Causality. According to [5], K-parser deals with causal relationships by detecting discourse markers such as "because" and then extracting the relation *caused_by*. However, we were not able to reproduce this behavior. Let us consider the sentence "Trump wants to ban Twitter because of the criticism he is receiving". K-parser considered the causal part separately, without making the connection between the two facts. Nonetheless, it linked the "wants" and "ban" events with the "objective" relationship. K-parser adds this binding when we have an event as an argument of another event.

has(Trump, is_subclass_of, person)
has(Trump, semantic_role, lawgiver)
has(Trump, semantic_role, wanter)
has(wants, agent, Trump)
has(ban, agent, Trump)
has(ban, objective, wants)
has(ban, is_subclass_of, social)
has(ban, recipient, twitter)
has(receive, is_subclass_of, possession)
has(receiving, supporting_verb, is)
has(criticism, is_subclass_of, communication)

Anaphoras are left unlinked. Consider "Barack Obama is an American politician. He was elected as the 44th president of the United States". The system returns a long list of instance-of and subclass-of facts, where the crucial ones leave the pronoun untouched:

has(He, is_subclass_of, person)
has(He, semantic_role, candidate)
has(Barack Obama-1, trait, American-4)

Anteriority. The system generally detects the chronological order of the events in an input sentence and extracts relations such as *next_event* and *previous_event*. However, whether this works or not depends on the time conjunction. For example, for "Once the votes are closed, the new president will be known", we obtained two separate events. For "Monica took her driver's license before buying a car", we obtain the correct chaining of events (as shown by this excerpt):

has(took, agent, Monica)
has(took, next_event, buying)

Contrast. K-parser does not have a special treatment for contrasting events. It treats each fact independently.
N-ary relations. As a semantic parser, K-parser can easily attach additional participants to an event. Furthermore, it enriches the graph with additional relations based on its ontology, identifying, e.g., instances of *person, location, place* etc.
Inter-sentence relations. Much like FRED, K-parser has difficulties with longer sentences. It also cannot link facts across different sentences.
Summary. K-Parser is a semantic parser that provides semantic annotations in addition to extracting the facts from the input sentence. It explicitly flags negated verbs, but it does not resolve anaphoras, it does not make the distinction between a fact and a belief, and it cannot deal with relationships across sentences.

4. Analysis

As we have seen, modern IE systems can cover much more than simple triples. Still, some desiderata remain open:
Belief. With this kind of sentences, we need to retain the context in order not to affirm the belief statement as a fact. Open IE 5 handles this category of sentences very well, by explicitly creating a context object. MinIE, too, deals very well with this kind of sentences, by explicitly attributing the hypothesis to the subject. FRED, too can easily create nested statements in its graph output. The other systems erase the distinction between a belief and a fact. This is, of course, problematic for downstream applications.
Negation. ClausIE and Open IE 5 just use the negated verb as is. MinIE, K-Parser, and FRED, in contrast, are able to specifically identify the negative polarity of the sentence. This is useful, e.g., for querying: The user can ask for all negative statements, or for all statements with a certain predicate (finding also the negative ones).

We can also imagine applications for reasoning (where negative statements become counter-evidence for a hypothesis), or for machine learning (where negative sentences can serve as negative training examples).

Causality and contrast are more challenging. Only FRED can see such a link between two events – and only by copying the conjunction verbatim into the graph.

What we would want instead is an explicit causal relationship. For example, for the sentence "Trump wants to ban Twitter because of the criticism he is receiving", we would expect:

F1: "Trump", "wants to ban", "twitter"
F2: "Trump", "is receiving", "criticism"
F3: F1, "caused by", F2

Such a representation would allow querying for the causes of an event, or for identifying transitive chains of causality. Possible knowledge representation formalisms for this type of facts are studied in [9].

Anteriority could be treated similarly to causality. Currently, only Open IE 5 and FRED can make such a relationship between two events explicit: Open IE 5 by adding a temporal marker, and FRED by a link between two event nodes.

Again, in the ideal case, we would see something like the following (for our example sentence "Monica took her driver's license before buying a car"):

F1: "Monica", "took", "driver's license"
F2: "Monica", "bought", "a car"
F3: F1, "before", F2

Such a representation would allow for some temporal reasoning, establishing, e.g., transitive anteriority between two events that are not directly linked.

N-ary relations. All the systems we have studied can deal with *n*-ary relations.

Anaphoras. In the ideal case, a system would replace the pronoun by its referent. For example, consider "Barack Obama is an American politician. He was elected as the 44th president of the United States". We would expect:

"Barack Obama ", "is", " American politician"
"Barack Obama", "was elected as 44th president of", " US"

However, none of the systems we have studied links a personal pronoun to its referent. This is a pity, because anaphoras are quite common in everyday written language (since they avoid repeating the subject of a sentence). When the anaphoras are not linked, the resulting triples are useless for applications such as querying or reasoning. If they could be linked, an important additional source of information could be tapped.

Inter-sentence relations. None of the systems we have seen can establish links across sentences. This is an important shortcoming, since complex information is often split across several statements in order to avoid lengthy sentences. For example, not all events in a biography are necessarily narrated in chronological order (Wikipedia biographies, e.g., usually first discuss the public life of a person and then the personal life). Here, anteriority markers such as "After that" provide an important clue. However, the IE systems we have studied would completely lose the temporal order of events. The same goes for more complex chains of reasoning, where one sentence gives the reason for the preceding one. These links, likewise, are lost.

5. Conclusion

In this survey, we have studied how state-of-the-art IE systems deal with complex information that does not fit neatly into a single triple. While most systems can easily deal with *n*-ary relationships, none of them can perform anaphora resolution or see relationships beyond single sentences.

For future work, this survey can be extended by studying more information extraction systems:

StuffIE [17] can extract events from text and causal links between them. Unfortunately, we were unable to run it or to reach out to the developers.

Nestie [18] can extract nested phrases such as beliefs or conditions. Unfortunately, the system does not seem to be available online.

Pikes [19] is a semantic parser that can resolve anaphoras.

Graphene [20] is a semantic parser that focuses on n-ary predicate-argument structures.

Our survey can also be extended to more dimensions of complex information:

Conditions say that a statement is true if another statement is true.

Sentiments attribute a personal valuation to an event or an object.

We hope that these analyses will help to move

information extraction towards a better understanding of human language.

Acknowledgements. This work was partially funded by the grant ANR-20-CHIA-0012-01 ("NoRDF").

References

[1] G. Weikum, J. Hoffart, F. M. Suchanek, Knowledge harvesting: Achievements and challenges, in: LNCS, 2019.

[2] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: EMNLP, 2017.

[3] T. Plattner, D. Orel, O. Steiner, Flexible data scraping, multi-language indexing, entity extraction and taxonomies: Tadam, a swiss tool to deal with huge amounts of unstructured data, in: Computational Journalism Symposium, 2016.

[4] A. Gangemi, V. Presutti, D. Reforgiato Recupero, A. G. Nuzzolese, F. Draicchio, M. Mongiovì, Semantic web machine reading with fred, Semantic Web 8 (2017).

[5] A. Sharma, N. H. Vo, S. Aditya, C. Baral, Towards addressing the winograd schema challenge—building and using a semantic parser and a knowledge hunting module, in: IJCAI, 2015.

[6] L. Del Corro, R. Gemulla, Clausie: clause-based open information extraction, in: WWW, 2013.

[7] K. Gashteovski, R. Gemulla, L. d. Corro, Minie: minimizing facts in open information extraction, in: EMNLP, 2017.

[8] M. Schmitz, S. Soderland, R. Bart, O. Etzioni, et al., Open language learning for information extraction, in: EMNLP, 2012.

[9] F. M. Suchanek, The Need to Move Beyond Triples, in: Text2Story workshop, 2020.

[10] C.-H. Chang, M. Kayed, M. R. Girgis, K. F. Shaalan, A survey of web information extraction systems, TKDE 18 (2006).

[11] J. L. Martinez-Rodriguez, A. Hogan, I. Lopez-Arevalo, Information extraction meets the semantic web: a survey, Semantic Web (2020).

[12] S. Sarawagi, Information extraction, Now Publishers Inc, 2008.

[13] C. Niklaus, M. Cetto, A. Freitas, S. Handschuh, A survey on open information extraction, arXiv preprint arXiv:1806.05599 (2018).

[14] C. F. Baker, C. J. Fillmore, J. B. Lowe, The berkeley framenet project, in: ACL, 1998.

[15] O. Etzioni, M. Banko, S. Soderland, D. S. Weld, Open information extraction from the web, Comm. ACM 51 (2008).

[16] A. Fader, S. Soderland, O. Etzioni, Identifying relations for open information extraction, in: EMNLP, 2011.

[17] R. E. Prasojo, M. Kacimi, W. Nutt, Stuffie: Semantic tagging of unlabeled facets using fine-grained information extraction, in: CIKM, 2018.

[18] N. Bhutani, H. Jagadish, D. Radev, Nested propositions in open information extraction, in: EMNLP, 2016.

[19] F. Corcoglioniti, M. Rospocher, A. Palmero Aprosio, Extracting knowledge from text with pikes, in: ISWC, 2015.

[20] M. Cetto, C. Niklaus, A. Freitas, S. Handschuh, Graphene: Semantically-linked propositions in open information extraction, in: ACL, 2018.