An Evaluation of Information Extraction Tools for Identifying Health Claims in News Headlines

Shi Yuan
School of Economics and Management
Beihang University
Beijing, China
ystone1025@buaa.edu.cn

Bei Yu
School of Information Studies
Syracuse University
Syracuse, NY, USA
byu@syr.edu

Abstract

This study evaluates the performance of four information extraction tools (extractors) on identifying health claims in health news headlines. A health claim is defined as a triplet: IV (what is being manipulated), DV (what is being measured) and their relation. Tools that can identify health claims provide the foundation for evaluating the accuracy of these claims against authoritative resources. The evaluation result shows that 26% headlines do not include health claims, and all extractors face difficulty separating them from the rest. For those with health claims, OPENIE-5.0 performed the best with F-measure at 0.6 level for extracting "IV-relation-DV". However, the characteristic linguistic structures in health news headlines, such as incomplete sentences and non-verb relations, pose particular challenge to existing tools.

1 Introduction

Mass media is a major source of information about health-related research, policies, and business. On average, four in ten American adults reported following health news stories closely. Thus, the quality of health news plays an important role in public understanding of health science (Brodie et al., 2003). However, inaccuracy in health news has raised concerns among scientists, journalists, and the general public. The lack of training has been identified as the main cause of low-quality health news stories (Sarri et al., 1998; Voss, 2002).

This study aims for using NLP techniques to identify inaccuracy in health news reporting. The resulting tools may be used for monitoring the quality of health news and providing training examples for journalists and readers. To achieve this ultimate goal, we propose a two-step solution: first extract the health claims in news stories and then verify these claims against reliable sources, such as original research publications. In this study we focus on the first step to extract health claims from news headlines, because headlines serve the role of attracting readers to click and read the whole story (NPR, 2018), and thus the inaccuracies in headlines may be more consequential than those in the main stories.

FDA (2009) defined a health claim as "the relationship between a substance and a disease or health-related condition". Sumner et al. (2014) defined it more broadly as a triplet of three elements: Independent Variable (IV), Dependent Variable (DV), and their relation. IV is defined as what is being manipulated, DV as what is being measured, and relation as the words that describe the link between IV and DV. They have applied this definition for manually examining exaggerations in health-related claims in science publications, press releases and news articles. Therefore, we adopt this definition to represent health claims. For example, in the headline "*Drug suppresses spread of breast cancer caused by stem-like cells*", the "IV-relation-DV" is "*drug; suppresses; spread of breast cancer*".

Entity-relation extraction has been a fundamental task in the area of information extraction. Some general-purpose tools have been developed to extract entities and relations, such as OPENIE-5.0 (Mausam, 2016), OLLIE (Mausam et al., 2012) and REVERB (Fader et al., 2011; Etzioni et al., 2011). Some tools were developed for specific domains, such as SemRep for extracting relations in biomedical publications (Rindflesch and Marcelo, 2003). However, the health news headlines sometimes have

Proceedings of the Workshop on Events and Stories in the News, pages 34–43
Santa Fe, New Mexico, USA, August 20, 2018.

unique grammatical structures compared to regular sentences with "subject-predicate-object" structures. For example, the noun "*prevalence*" describes the relation in the headline "*Prevalence of estrogen receptor mutations in patients with metastatic breast cancer*". Therefore, existing information extraction tools may encounter challenges in identifying "IV-relation-DV" from health news headlines.

In this paper, we evaluate popular information extraction tools for identifying health claims in the format of "IV-relation-DV" triples in health news headlines. The result is expected to shed light on the directions for improving the existing tools. The rest of paper is organized as follows. Section 2 reviews the state-of-the-art information extraction tools. Section 3 describes the construction of the benchmark data set. Section 4 presents the evaluation methods and results. Section 5 discusses the challenges that current tools face and offers suggestions for adaptation. Section 6 concludes the paper.

2 Related Work on Entity-relation Extraction

Studies in open information extraction aim for recognizing general-purpose "subject-relation-object" triplets from text. Based on the order of entity and relation identification, existing methods can be summarized into three types: identify entities first and then relations, identify relations first and then entities, and simultaneously identify both.

The first type of studies identifies entities first and then specifies some patterns to extract relations. REXTOR (Katz and Lin, 2000) is an early system that uses grammar rules for entity and relation extraction. Later, WOE^{Parse} (Wu and Weld, 2010), trained on Wikipedia articles, used dependency route patterns to decide whether two entities have relations. Besides, Relnoun (Pal and Mausam, 2016) was developed to extract relations from compound noun phrases, such as "*Collins; be director of; NIH*" in "*Collins, the director of NIH*". The designs of these methods raise some questions regarding their adaptability to our task. For REXTOR, the grammar rules would link entities to several types of relations that do not apply to our task, such as "is subject of". WOE^{Parse} would match entities to what appear in Wikipedia; however, IVs and DVs in our task may be new and thus not yet included in Wikipedia. For Relnoun, health claims in our task are not usually expressed in compound noun phrases.

Different from the first method, some studies prefer recognizing relations first. For example, REVERB (Fader et al., 2011; Etzioni et al., 2011) first extracts the longest word sequence that satisfies certain syntactic and lexical constraints as a relation. It then searches for the entities as two noun phrases that are nearest to the relation phrase, one on left, one on right. Experiment results have shown that REVERB has 30% higher AUC than WOE^{Parse} based on the precision-recall curve (Etzioni et al., 2011). Similar to REVERB, SRLIE (Christensen et al., 2011) is also a verb-centric system. It first identifies all verbs and their modifiers, and then extracts the verbs with at least two modifiers as relations. However, REVERB may mistakenly identify modifiers to IVs and DVs as entities when they are closer to the relation verbs. In Comparison, SRLIE tends to ignore the modifiers to IVs and DVs.

The third method is designed for simultaneously identifying entities and relations. This method usually depends on "subject-relation-object" patterns. For example, OLLIE (Mausam et al., 2012) uses dependency parsing patterns to identify triplets. Before OLLIE, TEXTRUNNER (Banko et al., 2007), a CRF-based system, uses part-of-speech tags for triplet identification. Empirical results showed that OLLIE performed better than REVERB based on precision-yield curve (Mausam et al., 2012), and REVERB better than TEXTRUNNER based on precision-recall curve (Etzioni et al., 2011). Furthermore, considering the need for identifying numerical relations, BONIE (Saha et al., 2017) applied numerical dependency patterns to extract triplets that a number or a quantity-unit phrase, such as "*Hong Kong; has labour force of; 3.5 million*" in "*Hong Kong's labour force is 3.5 million*".

Based on OLLIE, a new generation extractor named OPENIE-5.0 has been developed (Mausam, 2016). It combined SRLIE, Relnoun, BONIE and ListExtraction (Extraction from conjunctive sentences). Compared by precision-yield curve, OPENIE-5.0 is better than both OLLIE and REVERB (Mausam, 2016).

Overall, the aforementioned extractors rely on structural information in complete sentences to identify triplets. However, news headlines are often times not complete sentences. Therefore, whether the applicability of these extractors remain an open question.

In addition, because health news involves many biomedical concepts, we also reviewed information extractors in the biomedicine domain. SemRep (Rindflesch and Marcelo, 2003) is the state-of-the-art tool to identify semantic predications from biomedical text. It is a widely-used, rule-based system. The rules were derived from phrase structures (e.g., appositive structures). Furthermore, SemRep also relies on UMLS, a biomedical knowledge database, to identify concepts and relations. Besides SemRep, Dey et al. (2007) also proposed a system, which summarized PubMed articles by combining entity-relation structures into a network. The method for entity-relation structure identification depends on dependency relation rules. In comparison, SemRep suits our task better because the goal of Dey et al. (2007) is for text summarization instead of information extraction.

Based on the above review of the strength and weakness of the information extraction tools, we choose to evaluate four tools that best suit our task for extracting "IV-relation-DV" triplets in health news headlines: two representative systems of different methods (REVERB and OLLIE), a combination system (OPENIE-5.0) and a specific tool tailored to biomedicine (SemRep).

3 Benchmark Dataset Construction

We created a benchmark data set for evaluating the information extraction tools. This section describes the process of data collection, annotation, and validation.

3.1 Data Collection

`ScienceDaily.com` is a large website that aggregates science news. We collected all health news from `ScienceDaily.com` in 2016 and 2017, and selected all news articles with headlines including two common diseases "breast cancer" and "diabetes". The final collection contains 564 news articles, including 212 news headlines on breast cancer and 352 on diabetes. Those news headlines have been all annotated health claims manually.

3.2 Data Annotation Schema

We developed an annotation schema that includes three types of health claims: the first type does not describe a health claim, the second type describes a health claim between an IV and a DV, and the third type describes a health claim among multiple quasi-IVs (Sumner et al., 2014). Specifically, we define the annotation schema as:

- Health Claim or Not: Label a headline as "1" if it describes a health claim, otherwise label "0". For example, the headline "*Diabetics who use verapamil have lower glucose levels, data show*" is labeled as "1", but the headline "*Better breast cancer drugs?*" is labeled as "0". Sometimes health claims are phrased as questions in headlines, such as "*Can mindful eating help lower risk of type 2 diabetes, cardiovascular disease?*" For such cases, we also label them as "1".

- IV: What is being manipulated, e.g., "*Fasting-mimicking diet*" is the IV in the headline "*Fasting-mimicking diet may reverse diabetes*". The annotated IV should include all relevant words, including the modifiers. For example, the IV of "*Teen girls with a family history of breast cancer do not experience increased depression or anxiety*" is "*Teen girls with a family history of breast cancer*". Label "0" if no IV is found.

- DV: What is being measured, e.g., "*diabetes*" in in the headline "*Fasting-mimicking diet may reverse diabetes*". Label "0" if no DV is found. Similar to the IV annotation, the annotated DV should include all relevant modifiers. For example, the DV of "*Smoking can hamper common treatment for breast cancer*" is "*common treatment for breast cancer*".

- Relation: The statement of relation between IV and DV. The annotated relation should include all modifiers, like modal verbs (e.g., "*can*"), negative words (e.g., "*not*"), preposition combinations (e.g., "*associated with*") and verb combinations (e.g., "*help reduce*"). For example, the relation is "*found to switch*" in "*Breast cancer cells found to switch molecular characteristics*".

- Multiple IVs: Sometimes multiple quasi-IVs were mentioned if they are correlated (Sumner et al., 2014). In these cases, they are described in the same phrase, and thus impossible to separate as two independent phrases. For example, "*heart hormones, obesity, and diabetes*" in the headline "*New links between heart hormones, obesity, and diabetes*".

3.3 Inter-coder Agreement

We randomly chose 100 news headlines to evaluate inter-coder reliability. Two annotators annotated them separately. Inter-coder agreement was then calculated using Cohen's Kappa. We first evaluate the agreement on whether a headline describes a health claim (see Table 1). The Cohen's Kappa for this annotation task is 0.71, indicating substantial agreement. The main disagreement is on headlines with non-verb words to express relations, such as "*behind*" in "*Identifying a genetic mutation behind sporadic Parkinson's disease*", which was neglected occasionally.

		Annotator B	
		No Health Claim	**Health Claim**
Annotator A	**No Health Claim**	14	9
	Health Claim	0	77

Table 1: Confusion matrix from whether a headline describe a relation.

We then compared inter-coder agreement on IV, DV, and relation annotations on the 77 headlines with health claims identified by both annotators. Since these annotations are text snippets rather than categories, we convert the original annotations to five categories: "IV", "DV", "Relation", "Multiple IVs" and "No Annotation", and then examine each text snippet that has been annotated by either annotator, assigning it to the annotator's chosen category. Since annotations from two annotators may not be totally the same, we consider two annotations are the same if they share the main keywords or phrases. For example, consider a headline "*Breast, ovarian cancer may have similar origins, study finds*". One person annotated it as "IV-relation-DV" structure, "*Breast, ovarian cancer; may have; similar origins*" while the other annotated as "relation-Multiple IVs" structure, "*may have similar origins; Breast, ovarian cancer*". The two annotations correspond to each other as "IV" vs. "Multiple IVs", "Relation" vs. "Relation" and "DV" vs. "Relation". The confusion matrix was then generated accordingly (Table 2), and Cohen's Kappa is 0.89.

There are mainly two types of disagreement on IV, DV, and relation annotations. One is how to distinguish "IV-relation-DV" and "relation-Multiple IVs" structures on headlines with multiple IVs or DVs, such as "*Breast, ovarian cancer may have similar origins, study finds*". The other type of disagreements is whether a preposition phrase should be "DV" or not. For example, as for the headline "*A novel cancer immunotherapy shows early promise in preclinical studies*", one annotator annotated "*preclinical studies*" as "DV", but the other thought "*preclinical studies*" an adverbial modifier of "*shows early promise in*" and no "DV" in that headline.

		Annotator B				
		IV	**Relation**	**DV**	**Multiple IVs**	**No Annotation**
Annotator A	**IV**	68	3	2	2	2
	Relation	0	75	1	0	1
	DV	0	2	71	1	3
	Multiple IVs	0	0	0	0	0
	No Annotation	0	0	0	0	0

Table 2: Confusion matrix from IV, DV and relation annotations.

3.4 The Annotated Dataset

After the inter-coder reliability check, the disagreements were resolved through discussion. Then one annotator annotated the remaining news headlines. Among the 564 headlines, 416 (74%) describe health claims and 148 (26%) do not. Each of the 416 headlines with health claims was annotated as one "IV-relation-DV" triplet with a few exceptions. Five headlines were annotated with "Multiple IVs" (e.g., "*Little to no association between butter consumption, chronic disease or total mortality*"); two described more than one "IV-relation-DV" triplet (e.g., "*Epigenetic modification increases susceptibility to obesity and predicts fatty liver*"). We have also identified six types of linguistic structures in health news headlines that might confuse the extractors:

- Non-verb relation: such as "*New potential treatment for cancer metastasis identified*".
- Relation with modal verb: such as "*Smoking can hamper common treatment for breast cancer*".
- IVs and DVs with prepositional phrases: such as "*Sugars in Western diets*" in "*Sugars in Western diets increase risk for breast cancer tumors and metastasis*".
- Multiple IVs or DVs in parallel phrases: parallel phrases mean to contain more than one IVs or DVs, such as "*breast, ovarian cancer*" in "*Breast, ovarian cancer may have similar origins, study finds*".
- Headlines containing both reporting verb and another verb to describe the relation: such as "*find*" and "*treat*" in "*Scientists find 'outlier' enzymes, potential new targets to treat diabetes, inflammation*".
- Headlines as incomplete sentences: such as "*Sugar-sweetened drinks linked to increased visceral fat*".

The 416 headlines with health claims include 113 with incomplete sentence structure (27%), 39 with reporting verbs (9%), 39 non-verb relations (9%), 108 with modal verbs (26%), 107 with prepositional phrases in IVs (26%), 182 with prepositional phrases in DVs (44%), 18 with parallel phrases in IVs (4%), 42 with parallel phrases in DVs (10%). One headline may include multiple characteristics.

As a robustness check, we further compared the linguistic characteristics of the headlines about two diseases: "breast cancer" and "diabetes", and found no significant difference (see Figure 1). Therefore, we consider the linguistic characteristics of health news headlines independent of disease types.

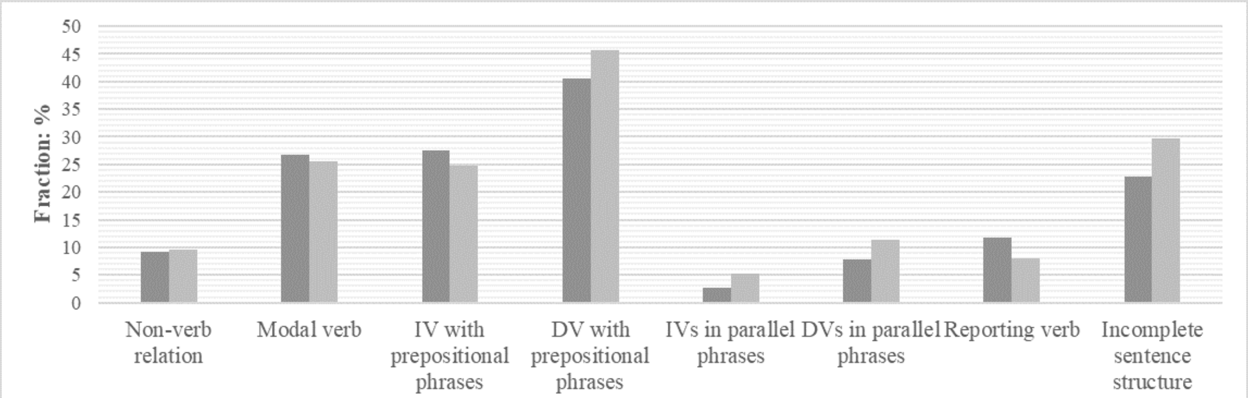

Figure 1: Distribution of different linguistic structures on breast cancer and diabetes.

4 Experiment Method and Result

Four systems were compared in terms of their performance in extracting "IV-relation-DV" from health news headlines. Section 4.1 describes the evaluation method. Section 4.2 evaluates performance on identifying headlines without health claims, and Section 4.3 on headlines with health claims. Section 4.4 evaluates performance on cases with special linguistic structures.

4.1 Evaluation Method

We chose two methods for this evaluation: the first one calculates the precision, recall and F-measure by manually comparing the machine annotations against the gold standard; the second method automatically calculates the BLEU scores between machine annotations and the gold standard.

Precision, Recall and F-measure are traditional evaluation methods in information retrieval. For information extraction task, Fader et al. (2011) and Etzioni et al. (2011) defined precision as the fraction of returned extractions that are correct, and recall as the fraction of correct extractions in the total corpus. For each extracted triplet, we manually check whether it is correct or not. The correct extraction is defined as keywords or phrases match in IV, DV and relation.

For robustness check we choose the second evaluation method as the BLEU score, which can be automatically calculated. As a popular measure in machine translation, it evaluates the similarity between the machine translations and the gold standard. The BLEU score divides each translated sentence into several n-grams and compares differences between a machine translation and a professional human translation based on those n-grams (Papineni et al., 2002).

In our task, we consider each manually-annotated "IV-relation-DV" triplet as a reference translation, and each machine-extracted triplet as a candidate translation. For each extractor, we will calculate a BLEU score. The higher BLEU score is, the better performance an extractor would achieve. The maximum number of n-grams varies from 2 to 4 and the weights for each n-gram are equal. In addition, we apply "Add-one Smoothing" technique proposed in Lin and Och (2004) to avoid the BLEU score being 0 when n grows bigger.

4.2 Headlines without Health Claim

In our benchmark data set, 26% headlines do not contain health claims. Correct extractions should return no triplets for such headlines. To evaluate the extractors on this task, we define precision as the correct results with no triplets among all results with no triplets, and recall as the correct results with no triplets among all headlines without claims. Table 3 shows the confusion matrix from each extractor and Table 4 shows the precision, recall and F-measure. The result shows that no extractors performed well in this task with all F-measures below 0.5 with little variation. Furthermore, OLLIE and OPENIE-5.0 had better precisions, but REVERB and SemRep had better recalls.

REVERB		Headlines without claims	Headlines with claims
	Results with no triplets	114	208
	Results with triplets	34	208
OLLIE		Headlines without claims	Headlines with claims
	Results with no triplets	77	107
	Results with triplets	71	309
OPENIE-5.0		Headlines without claims	Headlines with claims
	Results with no triplets	60	60
	Results with triplets	88	356
SemRep		Headlines without claims	Headlines with claims
	Results with no triplets	115	262
	Results with triplets	33	154

Table 3: Confusion matrix from different tools.

Information Extractor	Precision	Recall	F-measure
REVERB	.35	.77	**.48**
OLLIE	.42	.52	.46
OPENIE-5.0	**.50**	.41	.45
SemRep	.31	**.78**	.44

Table 4: Precision, Recall and F-measure for different tools on headlines without health claim.

An analysis of the false positive extractions shows that the main problem is the broad definition of relation in the three general-purpose tools, and thus many verbs that do not describe health claims were still identified as relations. For example, consider a news headline without health claim, "*New study explores concerns of African American breast cancer survivors*", the triplet is "*New study; explores; concerns of African American breast cancer survivors*" from REVERB, OLLIE, OPENIE-5.0. In contrast, SemRep is stricter on the definition of relations, which results in higher recall, but at the same time low precision with many headlines with health claims not identified as well.

4.3 Headlines with Health Claim

In this section, we evaluate the extractors' performance on headlines with health claims. The manual evaluation (Table 5) shows that OPENIE-5.0 ranked highest in F-measure at .62, followed by OLLIE at .53, REVERB at .41 and SemRep at .13. In Table 6, OPENIE-5.0 also ranked highest in BLEU score, followed by REVERB and OLLIE, regardless of the maximum number of n-grams. Overall, OPENIE-5.0 is the best tool for extracting "IV-relation-DV" in headlines with health claims, leaving some room for improvement. The results also show that the SemRep, the tool dedicated to relation extraction in biomedical literature, does not generalize well to health news. In addition, REVERB and OLLIE achieved similar precision level to OPENIE-5.0, but their recalls are relatively lower.

Information Extractor	Precision	Recall	F-measure
REVERB	.61	.31	.41
OLLIE	.62	.46	.53
OPENIE-5.0	**.67**	**.57**	**.62**
SemRep	.23	.08	.13

Table 5: Precision, Recall and F-measure for different tools on headlines with health claim.

Information Extractor	BLEU Scores (N=2)	BLEU Scores (N=3)	BLEU Scores (N=4)
REVERB	.66	.66	.65
OLLIE	.61	.58	.54
OPENIE-5.0	**.74**	**.71**	**.69**
SemRep	.17	.13	.10

Table 6: BLEU scores for different tools on headlines with health claim.

Table 3 has shown the number of false negative extractions from all tools. SemRep has 262, the highest number, followed by REVERB, 208, OLLIE, 107, and OPENIE-5.0, 60. Among those headlines, 24 headlines are the most challenging, because none of the extractors was able to identify the triplets. 71% of them (17 out of 24) are incomplete sentences. Since OPENIE-5.0 is the best performing system, we examined the missing triplets in its output and found 58% (35 out of 60) are incomplete sentences. Therefore, incomplete sentences are particularly challenging for the extractors.

Our benchmark data set includes only five headlines with multiple IVs and two headlines with multiple relations. Because these are likely difficult cases, we particularly checked the extractors' performance on these headlines. For the five headlines annotated with "relation-Multiple IVs" structure, REVERB and SemRep both return no triplets. OPENIE-5.0 returns two triplets and OLLIE returns three, but none of those triplets were correct. For headlines with more than one "IV-relation-DV" triplets, OPEINIE-5.0 and REVERB return one correct triplet "*Epigenetic modification; predicts; fatty liver*" of "*Epigenetic modification increases susceptibility to obesity and predicts fatty liver*", while others return no triplet or wrong triplets.

4.4 Impact of Linguistic Structures in Headlines

We then further examined the impact of the specific linguistic structures described in Section 3.4 on individual extractors (Table 7). If ranking the task difficulty by the best F-measure for each linguistic type, identifying non-verb relation is the most challenging with best F-measure at 0.19 by SemRep. Second, identifying triplets in incomplete sentences is also challenging with best F-measure at 0.30 by OLLIE and 0.28 by OPENIE-5.0. The extractors performed slightly better on two tasks with best F-measure at 0.40 level: identifying multiple IVs or DVs in parallel phrases, and identifying actual verb relations when reporting verbs are used. The extractors performed best on the tasks of identify-ing verb relations (including modal verbs) and identifying prepositional phrases in IVs and DVs with best F-measures over 0.6.

Structure Type	**Information Extractor**	**Precision**	**Recall**	**F-measure**
Non-verb Relation	REVERB	0	0	0
	OLLIE	.10	.05	.07
	OPENIE-5.0	0	0	0
	SemRep	**.33**	**.13**	**.19**
Verb Relation	REVERB	.64	.34	.44
	OLLIE	.66	.50	.57
	OPENIE-5.0	**.71**	**.63**	**.67**
	SemRep	.22	.08	.12
Modal Verb	REVERB	.70	.48	.57
	OLLIE	.76	.65	.70
	OPENIE-5.0	**.88**	**.88**	**.88**
	SemRep	.34	.11	.17
Prepositional Phrase in IV	REVERB	.35	.18	.24
	OLLIE	.57	.46	.51
	OPENIE-5.0	**.65**	**.57**	**.61**
	SemRep	.10	.05	.06
Prepositional Phrase in DV	REVERB	.67	.35	.46
	OLLIE	.68	.55	.61
	OPENIE-5.0	**.72**	**.63**	**.67**
	SemRep	.19	.10	.13
Parallel phrases for Multiple IVs	REVERB	.25	.06	.09
	OLLIE	.29	.22	.25
	OPENIE-5.0	**.40**	**.33**	**.36**
	SemRep	.11	.06	.07
Parallel phrases for Multiple DVs	REVERB	.67	.19	.30
	OLLIE	.41	.31	.35
	OPENIE-5.0	**.46**	**.41**	**.43**
	SemRep	.16	.07	.10
Headlines with Reporting Verbs	REVERB	.36	.23	.28
	OLLIE	.23	.18	.20
	OPENIE-5.0	**.41**	**.38**	**.39**
	SemRep	.29	.10	.15
Incomplete Sentences	REVERB	.50	.09	.15
	OLLIE	**.40**	**.24**	**.30**
	OPENIE-5.0	.35	.24	.28
	SemRep	.29	.11	.16

Table 7: Performance on different linguistic structures.

5 Challenges to Individual Extractors

Based on the above results, we summarize the main challenges for each extractor and offer suggestions for improvement.

SemRep: SemRep outputs significantly fewer triplets than the other tools. It even missed the cases with clear verb structures, such as "*Smoking can hamper common treatment for breast cancer*". The restriction may be attributed to SemRep's strict definition on some entities and relations. For example, "associated with" is defined as the relation between a gene and a disease only (Kilicoglu et al., 2011). Therefore, loosening the definition on some entities and relations might be helpful.

REVERB: Since REVERB is a verb-based extractor, the first suggestion is to add some rules for processing headlines with non-verb relations. In addition, the ability to recognizing IV and DV in complicated noun phrases should also be improved. In such cases, REVERB missed the head nouns in complicated noun phrases. For example, the IV extracted by REVERB is "*humans*" in healine "*One of the most common viruses in humans may promote breast cancer development*".

OLLIE: Headlines with reporting verb bring a big problem to OLLIE. OLLIE tends to extract the reporting verbs as the relations while ignore the actual verb relations. For example, for the headline "*Scientists find 'outlier' enzymes, potential new targets to treat diabetes, inflammation*", OLLIE identified "*Scientists; find; 'outlier' enzymes*" as the triplet, but missed the actual health claim that the enzymes may be able to treat diabetes.

OPENIE-5.0: OPENIE-5.0 faces the challenges of non-verb relations and incomplete sentences. Especially, for headlines with "A linked to B" structure, OPENIE-5.0 can only identify triplets without DVs. For example, OPENIE-5.0 identified "*Sugar-sweetened drinks; linked;*" in "*Sugar-sweetened drinks linked to increased visceral fat*".

6 Conclusion

In this paper, we have created a benchmark data set, and used both manual and automated evaluation methods to compare the performance of four information extractors on identifying the health claims in health news headlines. Both methods reached consistent findings. Overall, 26% of health news headlines do not include health claims, and 74% do. The three general-purpose extractors (OPENIE-5.0, OLLIE, REVERB) performed better than the biomedicine-specific extractor (SemRep), probably because SemRep was developed for documents in academic writing, not popular science writing. Among those general-purpose extractors, OPENIE-5.0 has the best performance to extract "IV-relation-DV" triplets with F-measure at 0.6 level. However, some characteristic linguistic structures in health news headlines pose particular challenge to these extractors, especially on identifying non-verb relations and relations in incomplete sentences. With F-measure at 0.4 level, further improvement is needed for identifying multiple IVs or DVs in parallel phrases, or identifying actual verb relations when reporting verbs are around. The extractors can identify verb relations and prepositional phrases in IVs and DVs relatively well with F-measure at 0.6 level. In future work, we would like to develop new tools for identifying headlines without claims and enrich the current rule-based systems with rules tailored to the linguistic characteristics of health news headlines.

Acknowledgements

We would like to thank Dr. Jun Wang for providing the data. This research was partially supported by China Scholarship Council and Syracuse University.

References

Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead and Oren Etzioni. 2007. *Open information extraction from the web*. In IJCAI (Vol. 7, pp. 2670-2676).

Mollyann Brodie, Elizabeth C. Hamel, Drew E. Altman, Robert J. Blendon and John M. Benson. 2003. *Health news and the American public, 19962002*. Journal of Health Politics, Policy and Law, 28(5), 927-950.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2011. *An Analysis of Open Information Extraction Based on Semantic Role Labeling*. In Proceedings of the Sixth International Conference on Knowledge Capture (pp. 113120). New York, NY, USA: ACM.

Lipika Dey, Muhammad Abulaish, Jahiruddin and Gaurav Sharma. 2007. *Text Mining through Entity-Relationship Based Information Extraction*. In 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops (pp. 177180).

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam. 2011. *Open Information Extraction: The Second Generation*. In IJCAI (Vol. 11, pp. 3-10)

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. *Identifying Relations for Open Information Extraction*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 15351545). Stroudsburg, PA, USA: Association for Computational Linguistics.

FDA. 2009. *Guidance Documents & Regulatory Information by Topic - Guidance for Industry: Evidence-Based Review System for the Scientific Evaluation of Health Claims*. `https://www.fda.gov/Food/GuidanceRegulation/GuidanceDocumentsRegulatoryInformation/ucm073332.htm` (last access: May 10, 2018)

Boris Katz and Jimmy Lin. 2000. *REXTOR: A System for Generating Relations from Natural Language*. In ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval (pp. 6777). Hong Kong, China: Association for Computational Linguistics.

Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman and Thomas C Rindflesch. 2011. *Constructing a semantic predication gold standard from the biomedical literature*. BMC Bioinformatics, 12, 486.

Chin-Yew Lin and Franz Josef Och. 2004. *Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-bigram Statistics*. In Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics.

Mausam. 2016. *Open information extraction systems and downstream applications*. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (pp. 4074-4077). AAAI Press.

Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. *Open Language Learning for Information Extraction*. In Proceedings of the 2012 Joint Conference on Empirical Meth-ods in Natural Language Processing and Computational Natural Language Learning (pp. 523534). Stroudsburg, PA, USA: Association for Computational Linguistics.

NPR. 2018. *How to make great headlines*. `http://training.npr.org/digital/the-checklist-for-writing-good-headlines/` (last visited May 14th, 2018)

Harinder Pal and Mausam. 2016. *Demonyms and Compound Relational Nouns in Nominal Open IE*. In Proceedings of the 5th Workshop on Automated Knowledge Base Construction (pp. 3539). San Diego, CA: Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: A Method for Automatic Evaluation of Machine Translation*. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (pp. 311318). Stroudsburg, PA, USA: Association for Computational Linguistics.

Thomas C. Rindflesch and Marcelo Fiszman. 2003. *The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text*. Journal of Biomedical Informatics, 36(6), 462477.

Mary-Anne Saari, Candace Gibson and Andrew Osler. 1998. *Endangered species: science writers in the Canadian daily press*. Public Understanding of Science, 7(1), 61-81.

Swarnadeep Saha, Harinder Pal and Mausam. 2017. *Bootstrapping for Numerical Open IE*. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 317323). Vancouver, Canada: Association for Computational Linguistics.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aime Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy and Christopher D Chambers. 2014. *The association between exaggeration in health related science news and academic press releas-es: retrospective observational study*. BMJ, 349, g7015.

Melinda Voss. 2002. *Checking the Pulse: Midwestern Reporters' Opinions on Their Ability to Report Health Care News*. American Journal of Public Health, 92(7), 11581160.

Fei Wu and Daniel S. Weld. 2010. *Open Information Extraction Using Wikipedia*. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 118127). Stroudsburg, PA, USA: Association for Computational Linguistics.