Lang Resources & Evaluation
https://doi.org/10.1007/s10579-020-09491-8

PROJECT NOTES

Writer's uncertainty identification in scientific biomedical articles: a tool for automatic if-clause tagging

Paolo Omero[1] · Massimiliano Valotto[1] · Riccardo Bellana[1] · Ramona Bongelli[2] · Ilaria Riccioni[2] · Andrzej Zuczkowski[2] · Carlo Tasso[1]

Abstract In a previous study, we manually identified seven categories (verbs, non-verbs, modal verbs in the simple present, modal verbs in the conditional mood, if, uncertain questions, and epistemic future) of Uncertainty Markers (UMs) in a corpus of 80 articles from the British Medical Journal randomly sampled from a 167-year period (1840–2007). The UMs detected on the base of an *epistemic stance* approach were those referring only to the authors of the articles and only in the present. We also performed preliminary experiments to assess the manual annotated corpus and to establish a baseline for the UMs automatic detection. The results of the experiments showed that most UMs could be recognized with good accuracy, except for the if-category, which includes four subcategories: if-clauses in a narrow sense; if-less clauses; as if/as though; if and whether introducing embedded questions. The unsatisfactory results concerning the if-category were probably due to both its complexity and the inadequacy of the detection rules, which were only lexical, not grammatical. In the current article, we describe a different approach, which combines grammatical and syntactic rules. The performed experiments show that the identification of uncertainty in the if-category has been largely double improved compared to our previous results. The complex overall process of uncertainty detection can greatly profit from a hybrid approach which should combine supervised Machine learning techniques with a knowledge-based approach constituted by a rule-based inference engine devoted to the if-clause case and designed on the basis of the above mentioned epistemic stance approach.

Keywords Uncertainty markers · Epistemic stance · Scientific biomedical articles · Automatic if clause tagging · SVM approach · Rule-based approach

✉ Ramona Bongelli
ramona.bongelli@unimc.it

[1] University of Udine, Udine, Italy

[2] University of Macerata, Macerata, Italy

Published online: 11 June 2020

1 Introduction

The certainty or uncertainty of information communicated by biomedical scientific writers through a series of linguistic (both lexical and morphosyntactic) markers plays a significant role in determining whether that information will be translated into practice or not. For example, National Governments make decisions regarding their health policies on the basis of how certain or uncertain the results from biomedical research are communicated. On the same basis, the scientific community steers their own research and clinicians direct their practice.

1.1 Related work on uncertainty in scientific writing

Given the importance of certainty/uncertainty language in determining practical decision making, the field has received considerable attention in linguistics from scholars starting in the 90s (among the most known, Hyland 1994, 1995, 1998a, 1998b; Salager-Meyer 1994; Crompton 1997; Rubin 2007). Most of them mainly adopted a *top-down* approach, i.e., the uncertainty markers were initially extracted from grammar books and dictionaries and subsequently applied in their analysis, without referring to any explicit and comprehensive linguistic theory of certainty and uncertainty.

1.2 Related work on uncertainty in Natural Language Processing

Distinguishing certain (= factual) and uncertain (= speculative) information in texts is of crucial importance in information extraction (IE) as well. Indeed, the detection of certainty/uncertainty markers and their linguistic 'scope' (Quirk et al. 1985) has been receiving increasing attention in the Natural Language Processing (NLP) community (among the most known studies, Vincze et al. 2008; Kim et al. 2009; Özgür and Radev 2009; Agarwal and Yu 2010; Farkas et al. 2010; Szarvas et al. 2012; Zou et al. 2013; Zhou et al. 2011, 2015).

Although those studies analyse corpora annotated for uncertainty language, most of these annotations are not based on an explicit linguistic theory of certainty and uncertainty communication (they are mainly based on Hyland's 1994 list of uncertainty markers) and tend to be small in their number of full-text scientific articles. In addition, these studies lack a historical perspective to evaluate how uncertainty has evolved over time.

1.3 Previous study

Differently from the above mentioned studies, we analysed (Bongelli et al. 2012, 2014; Zuczkowski et al. 2016; Bongelli et al. 2019) a wide and diachronic corpus of biomedical full texts articles (80 articles from 1840 to 2007 randomly selected from the British Medical Journal)[1] on the base of an explicit linguistic

[1] The British Medical Journal is available from PubMed Central (http://www.ncbi.nlm.nih.gov/pmc/journals/3/, last accessed March 2011). Our corpus of articles is made up of 187,854 words.

theory of certainty and uncertainty language (Zuckowski et al. 2014a; Zuczkowski et al. 2017). The novelty of this approach, which combines, for the identification of the uncertainty markers, a top-down and a bottom-up method, is that it takes into account *only* the author's uncertainty in the here and now of communication and not the uncertainty of somebody else, mentioned in the article, as the above mentioned studies usually do (see Sect. 2).

After the manual annotation, we also performed preliminary experiments for the automatic detection of uncertainty markers (UMs), by using Machine-Learning techniques (Bongelli et al. 2012). The results of the experiments showed that most UMs could be recognized with good accuracy, except for the if-category.

In the current article, we describe a different automatic approach adopted to detect uncertainty in the if-category and constituted by grammatical and syntactic rules, based on our linguistic model of uncertainty. The new performed experiments show that the identification of uncertainty in the if-category has been largely double improved compared to our previous results.

In Sect. 2 our theory on certainty and uncertainty is outlined together with the results referring to our previous study. In Sect. 3, the software architecture for the automatic detection of the if-category is described. In the final Sect. 4, an evaluation of the proposed approach and plans for future activities are illustrated.

2 A theoretical perspective on certainty and uncertainty

2.1 The framework

The study of certainty and uncertainty in communication is related to the more general topics concerning *epistemicity* (e.g., Dendale and Tasmowski 2001; Nuyts 2001), *evidentiality* (e.g., Chafe and Nichols 1986; Willett 1988), *mitigation* (e.g., Caffi 2007), *hedging* (e.g., Lakoff 1973; Fraser 1980; Holmes 1984) and more specifically with *epistemic stance* (e.g., Ochs 1996; Kärkkäinen 2003; Stivers et al. 2011; Heritage 2012).

Our theory on certainty and uncertainty is a theory of *epistemic stance* since it focuses on the here and now of communication, i.e., on how speakers/writers communicate their certain or uncertain stance towards the information they are conveying (Zuckzkowski et al. 2017).

From this perspective, certainty and uncertainty can be defined in the following way: a piece of information is communicated as certain when, in the here and now of communication, the speaker/writer's commitment to its truth is at the maximum or high level; on the contrary, a piece of information is communicated as uncertain when, in the here and now of communication, the speaker/writer's commitment to its truth is at the minimum or low level.

When in a British Medical Journal (BMJ) article we read, for example:

(1) "It is certain that the cholera stools contain some poisonous materials…" (Johnson 1865).

the author communicates that it is certain for him that the piece of information p (= *the cholera stools contain some poisonous materials*) is true, i.e., he is saying that he evaluates p as true.

Uncertainty means that, if the author had written *Perhaps the cholera stools contain some poisonous materials*…, he would have said that he does not know whether p is true or false, therefore he would have communicated p as uncertain, i.e., he would have told the readers that he is not certain towards the truth of p.

2.1.1 The author's uncertainty in the here and now of communication

In written texts such as BMJ articles, UMs can refer either to the author's uncertainty or to somebody else's uncertainty. Both types of uncertainty can refer to the present or past or future.

As said above, an essential point in our study on BMJ is that we specifically aimed at identifying the writer's UMs referring to the here and now of his communication, i.e., at the time the article was being written.

We excluded from our analysis (1) the writer's UMs referring to the past or the future, and (2) the UMs of somebody else different from the author of the article.

Consider the following example:

(2) "I am not quite sure whether it was Dieffenbach or Jobert who first exposed the error of former operators" (Wells 1861).

In this example, the author communicates that he is currently uncertain about the information (…*it was Dieffenbach or Jobert who first exposed the error of former operators*) that follows the UM (*I'm not quite sure whether*…).

If the sentence were

(2a) I was not quite sure whether…

instead of "*I am not quite sure whether*…", the UM would again refer to the author's uncertainty but, unlike the original example, in the past and not in the present. It means that, in the here and now of communication, the author *remembers* that *there and then* (i.e., in the *past*) he was uncertain about the information. In other words, in the here and now, the author is communicating as *certain* an information concerning his past uncertainty.

If the sentence were

(2b) Doctor Collins is not/was not/will not be quite sure whether…

the UMs in the present, past or future would refer to Doctor Collins and not to the author. In other words, the author, in the here and now of communication, is communicating as *certain* an information referring to the uncertainty of someone different from himself.

As a consequence, our analysis had only detected uncertainty under the first case (the author's uncertainty in the present, example 2), and not the other two (the author's uncertainty in the past or future and somebody else's uncertainty in the present, past or future, examples 2a and 2b).

2.1.2 Reasons for our differentiated approach

To the best of our knowledge, no previous study has applied such distinction in the detection of UMs in biomedical field. Applying or not this distinction means to study two different types of issue and leads to different quantitative results.

When adopting our differentiated approach, only example (2) would be considered as uncertain; on the contrary, when adopting an undifferentiated approach, also examples (2a) and (2b) would be considered as uncertain. The former approach is specific, i.e., it takes only the UMs referring to the scientific writers in the present; the latter is generic, i.e., it takes any UM indiscriminately, thus mixing up in a senseless way anybody's (= writers' and non-writers') past, present, and future uncertainty.

The choice of one or the other approach differently affects the quantitative results concerning both the UMs and their linguist scope (Quirk et al. 1985). Indeed, in the latter case, the quantitative results would be wider since the undifferentiated approach considers not only the author's uncertainty in the present, but also in the past and future, as well as the present, past and future uncertainty of somebody else mentioned in the article (*Doctor Collins* in the examples 2a and 2b).

Of course, the differentiated approach, being more sophisticated, requires complex and specific rules both for the manual and the automatic detection.

2.1.3 Seven categories of UMs

On the basis of our theory on certainty and uncertainty, we manually identified seven categories of UMs, both lexical and morphosyntactic: verbs, non-verbs, modal verbs in the simple present, modal verbs in the conditional mood, if-category, uncertain questions, and epistemic future (see Fig. 1).

2.1.4 The if-category

As shown in Fig. 1, the if-category is divided in the following four sub-categories:

(1) *If-clauses* in a narrow sense: In English, there are different forms of *if-clauses*, i.e., *conditionals*. The different conditionals can be identified according to specific patterns of the tenses and moods exploited in the verbs of the protasis and of the apodosis.

The form called *zero conditional* occurs when *if* is accompanied by simple present in the protasis as well as simple present in the apodosis. This is the only situation in which we do not classify the if-clause as uncertain (i.e., the if-clause is not considered a UM) since in this case the *if* can be paraphrased with a temporal conjunction, for example 'when' and 'every time,' all of which communicate certainty. An example is:

Simple present + simple present

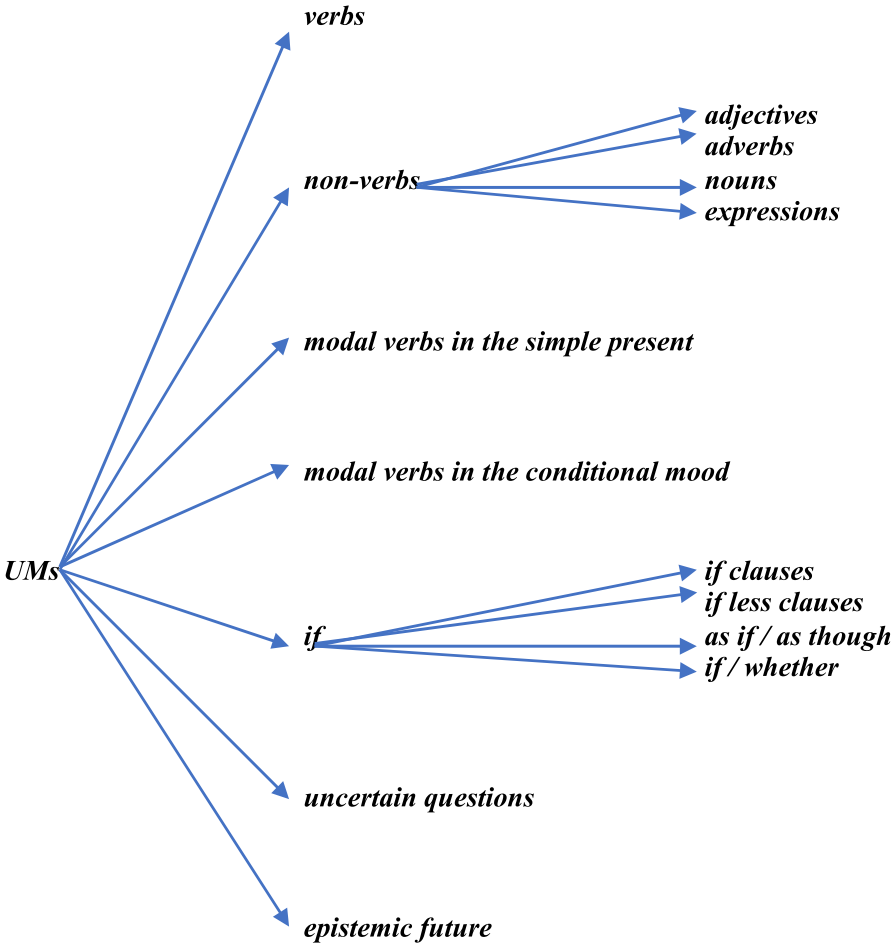

Fig. 1 UMs categories

(3) "if the discharges succeed each other very rapidly, the irides have an oscillatory motion" (Crichton-Browne 1875).

All other forms of if-clauses were considered as UMs, both (a) because the *if* in the protasis can be paraphrased as an *uncertain condition* on which the apodosis depends (see comment to example 4 below) and (b) because of the presence of a modal verb in the future tense (*will*) or in the conditional mood (*would*, *should*, etc.) in the apodosis. Unlike the simple present (zero conditional), modal verbs communicate possibility (uncertainty), not factuality (certainty) (Lyons 1977, chapter 17; Palmer 1986, chapter 5; Radden and Dirven 2007, chapter 10). Therefore, the rationale of the if-clauses classification takes into account the presence of both the *if* and the verb tense and mood, as in the following:

Simple present + simple future (*first conditional*), such as in

(4) "if it is early and properly performed, and if the after treatment is judicious, it will be even more successful than it has been in my hands" (Radford 1849).

In the protasis the *if* means that the author, in the here and now of communication, does not know whether *it is early and properly performed, and* whether *the after treatment is judicious*, i.e., he is uncertain about the realization of the conditions on which the apodosis depends. *If* these conditions occur, then *it will be even more successful than it has been in my hands*. An analogous comment can apply also to examples 5-8. In other cases, the present tense is not in the indicative mood, as it was in the previous example, but in the subjunctive mood, as in the following one.

(5) "If a vaccinated sheep be inoculated with anthrax within a few days of the operation, it will die of splenic fever" (Lister 1880).

Simple past + present conditional (*second conditional*), such as in

(6) "*If the association suggested by the upper part of the table were due merely to a bias in our method investigation, we would expect to see that bias operating to some extent in all, or nearly all, causes of death*" (Doll and Hill 1956).

Past perfect + perfect conditional (*third conditional*), having as an example

(7) "*These age rates for a smoking category were then applied to the corresponding U.K. population in 1951 to obtain the death rate at all ages that would have prevailed in the U.K. population if it had experienced the rates at specific ages of the particular smoking group*" (Doll and Hill 1956).

Simple present + present conditional (*mixed type conditional*), such as in

(8)"If we are content to record systolic pressures alone, unquestionably we should say that the man with a systolic pressure of 140 ran the graver risk" (Dally 1913).

(2) *If-less clauses.* We also tagged the implicit *if-clauses*, i.e., the constructions having in the protasis, instead of the explicit *if*, only the subject-verb inversion, such as

(9) "Had I regarded the systolic pressures alone, I should have said that the aortic case had the higher blood pressure, and that his arteries were in a condition of greater stress than those of the man with granular kidney" (Dally 1913).

In this example, the initial expression with the subject-verb inversion 'Had I regarded' is equivalent to 'If I had regarded'.

In English, the subject-verb inversion in the protasis of the if-less clauses can be made with the following three verbs, independently from the tense and mood of the verb in the apodosis (simple present, future, present conditional, past conditional, etc.):

had (past perfect), as in the above example (7);
should (present conditional):

(10) "should early thrombosis of the graft take place the circulation will not be reduced" (Horton 1956).

were (past simple):

(11) "Were it otherwise the pulmonary artery would be affected as often as the aorta" (Barr 1909).

(3) Comparative constructions introduced by *as if* and *as though* including the following examples:

(12) "If the stones look as though they may be difficult to remove endoscopically the surgeon should convert to open exploration of the bile duct" (Scott-Coombes and Thompson 1991).

(13) "The extracts behaved as if they contained noradrenaline" (Burn and Rand 1958).

In a statement of the form "p as if q" a comparison between the main clause p and a hypothetical clause q is established. In accordance with this and following Vaihinger's (1952) pioneer analysis of such propositions, they are interpreted as composed of a comparative clause and of an if-clause with understood apodosis (Zuczkowski et al. 2014b): *The extracts behaved as if they contained noradrenaline = The extracts behaved as* ***[they would]*** *if they contained noradrenaline = If the extracts contained noradrenaline,* ***they would*** *behave the way they did.*

For this reason, all clauses having *as* immediately before *if* or *though* were detected as UMs.

(4) *If* and *whether* introducing indirect uncertain questions (Zuczkowski et al. 2016; Bongelli et al. 2019). They were considered as UMs when referring to the writer's uncertainty in the here and now of communication, i.e., when both the verb that precedes the *if/whether* and the verb that follows it are not in the past, but in one of the following tense combinations:

Present + Present:

(14) "It is, in fact, doubtful whether the same proportionate degree of protection is likely to be conferred by vaccination in a community[…]" (Wilson 1947).

Present + Past:

(15) "It is not clear whether in the contact cases the controls were isolated in exactly the same way" (Wilson 1947).

Past + Present:

(16) *"In this study we have endeavoured to find out whether, after initial assessment in hospital, it is possible to maintain an adequate reduction in pressure on a long-term out-patient basis without undue side-effects"* (Lowther and Turner 1963).

Present + Future:

(17) "*It is never possible to predict before making the attempt whether or not one will be able to produce pneumothorax*" (Lucas 1915).

During the manual annotation, it was noted that, differently from the above four examples, in some sentences related to the combination Present + Present and Present + Past the clauses preceding and following the *if* or *whether* were inverted, i.e., the clause that usually precedes the *if* or *whether* was placed after the clause that generally follows it:

(18) "I usually suture the edge of the internal oblique down into the groove of Poupart's ligament as far as I am able. **Whether** this **is** much real use I **am** not **prepared** to say" (Davies 1913).

(19) "In 1946 a lumbar sympathectomy was performed. **Whether** or not this **has influenced** the good result **is** open to doubt" (Bourne 1955).

2.1.5 Quantitative results

According to the results of our manual annotation, the if-category includes 313 occurrences of UMs:

- 195 if-clauses;
- 24 if-less clauses;
- 7 as if/as though (= 6 as if + 1 as though);
- 87 if/whether.

2.2 Automatic annotation of UMs

As mentioned in the introduction, we performed preliminary experiments to assess the manually annotated corpus and establish a baseline for the automatic detection of UMs (Bongelli et al. 2012). The experiments were carried on using YamCha[2] that is a generic and customizable tool applied in different NLP tasks, such as POS tagging, Named Entity Recognition, base NP chunking, and Text Chunking. The

[2] YamCha is an open source text chunker. http://chasen.org/taku/software/yamcha/

Table 1 Results of preliminary experiments

UMs	Precision	Recall	F1
Modal verbs in the conditional mood	84.22	98.89	90.97
If	**26.42**	**93.99**	**41.24**
Modal verbs in the simple present	78.98	99.66	88.12
Non-verbs	77.09	79.31	78.19
Verbs	89.16	66.23	76.01
Overall	68.67	90.42	78.06

The results concerning the if category are in bold

automatic classification process was performed by means of a Machine Learning approach, namely, Support Vector Machines (SVMs).

The documents have been processed using TreeTagger[3] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/, a language-independent PoS tagger and the corpus annotation converted to IOB format.

The Machine Learning techniques used for the recognition of UMs obtained encouraging results for most of the categories of our classification. The results of these preliminary experiments indeed showed that most UMs were recognized with good accuracy (overall precision = 68.67%; recall = 90.42%; F1 score = 78.06%). As shown in Table 1 (Bongelli et al. 2012),[4] only the results concerning the *if*-category were substantially lower than those of the other categories (precision = 26.42; recall = 93.99; F1 score = 41.24).

The unsatisfactory results concerning the if-category were probably due to both its complexity (four sub-categories) and the inadequacy of the detection rules we had formulated at that time: only the lexical aspects were taken into consideration, not the grammatical ones.

3 The present study

In order to improve the results for the more sophisticated case of the if-category, we propose to deepen the approach to the automatic annotation process and to explicitly exploit grammatical and syntactic knowledge.

The rationale behind our proposal is the *higher level* of sophistication of the (cognitive) process required to detect if-clauses, as illustrated in the previous Sect. 2.1: further aspects have to be considered, such as positional features and precise recognition of verb tenses and moods. In other terms, our claim is that, in order to overcome the limitations of the approach presented in Bongelli et al. 2012, it is necessary to take into account more detailed linguistic rules such as those we are going to present in Sects. 3.3, 3.4, 3.5, 3.6 and 3.7.

As a consequence, we decided to develop a *knowledge-based* (more strictly speaking a *rule-based*) system, specifically devoted to if-clause identification. Such

[3] Treetagger is a language-independent PoS tagger.

[4] In the preliminary experiments, the *epistemic future category* was empty of UMs; the *uncertain question category* was not yet included.

system has to compute all the parameters used in the framework in order to recognize UMs.

From a more general point of view, our claim asserts that an empirical, supervised ML approach (such as SVM) may not perform adequately when the decision (classification) process is based on a sophisticated linguistic model. Instead, an approach based on explicit representation of linguistic knowledge may reach higher performance.

Software architecture and operation of the proposed rule-based system for if-clause detection is illustrated in the following, after a short review of related work.

3.1 Related work on the automatic detection of if-clauses

NLP studies have been mostly focused on uncertainty *lexical* markers rather than on more complex *grammatical* means, such as *if-clauses* (Thompson et al. 2011: 4). Only few works have investigated this topic in the biomedical field (see for example, Kilicoglu and Bergler 2008, 2010; Velldal et al. 2010; Velldal et al. 2012; Malhotra et al. 2013) as well as in non-biomedical contexts (see for example, Narayanan et al. (2009). Moreover, to the best of our knowledge, no previous study on the automatic detection of if-clauses has taken into account *if-less clauses* (i.e., our subcategory 2) and *as if/as though* (i.e., our subcategory 3). Finally, the systematic consideration of verb tenses and moods is another distinguishing feature of our approach.

3.2 Software architecture

The grammatical and syntactic knowledge that we exploit is represented in form of if–then rules, that we call *if-clause detection rules*. The if-part of the rule indicates the conditions to be satisfied in order to assign a specific if-category. Such conditions are expressed in terms of specific relevant words (if tokens), sequences of words, position, mood, and tense of verbs, according with the linguistic framework presented in Sect. 2.1. In other words, the automatic rules are based on the linguistic rules used for the manual detection.

The software architecture, presented in Fig. 2, is based on a pipeline of independent subsystems, each one devoted to a specific subprocess of the overall analysis.

The following subprocesses are included in the pipeline:

1. *Pre-processing*. Its aim is to normalize peculiar expressions, such as abbreviations or units of measurement, in order to avoid problems in the next processing steps.
2. *Quotation Identification*. Its aim is to identify and remove direct quotations because, even if they contained some UMs, they would not be related to the author of the article.
3. *Sentence Splitting*. Its aim is to split the whole document into sentences. Since in our research syntactic completeness is more important than the semantic one, we tried to obtain sentences as short as possible.

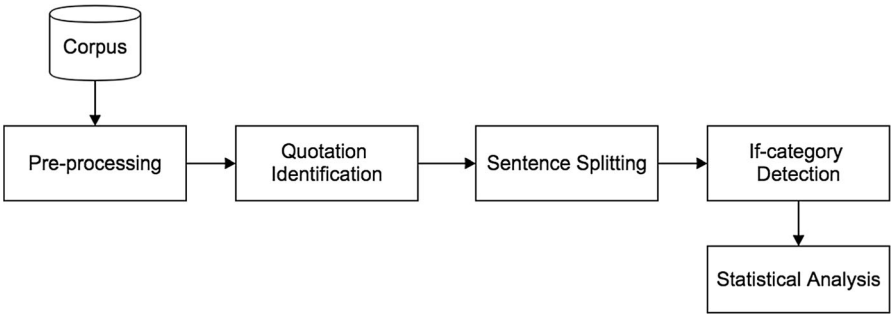

Fig. 2 Overall architecture and processing workflow

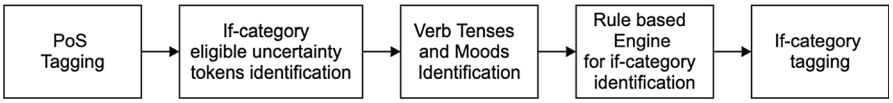

Fig. 3 Workflow of the if-category identification process

4. *If-category Detection*. It is the core module of the workflow. Its aim is to detect the *if-category* tokens, to determine which of them are UMs (see Sect. 2.1) and to tag them. In order to achieve this goal, the sentences are first processed by the Stanford PoS tagger.[5]
5. *Statistical Analysis*. Its aim is to compare the manually assigned tags with the automatically assigned ones and to evaluate the resulting precision, recall, and F1 score.

3.3 If-category Detection

The If-category Detection subsystem identifies the UMs belonging to the *if*-category by executing the following 5 steps:

1. PoS tagging each word of the sentence.
2. Identifying the presence of eligible uncertainty tokens ("if", "whether", "as if/as though", specific verbs, etc.) to be used in the next steps.
3. Identifying verb tenses and moods.
4. Matching the syntactic rules on the previously identified tokens and firing the appropriate ones.
5. Tagging tokens identified as UMs.

The *if-category* identification process is shown in Fig. 3. The if-clause detection rules exploited in the *Rule based Engine for if-category identification* task have

[5] https://nlp.stanford.edu/software/tagger.shtml.

Table 2 First subcategory rules (if-clauses in a narrow sense)

Type	Protasis	Apodosis
First conditional	Present simple	Future Simple
Second conditional	Past simple	Present conditional
Third conditional	Past perfect	Perfect conditional
Mixed Type	Present simple	Present conditional

been derived from the linguistic framework illustrated in Sect. 2.1: they are described in the next sections.

The activation (firing) of a rule includes two phases:

1. Evaluation of the if-part, in two steps: (*i*) each condition specified is computed and (*ii*) the overall satisfaction of the if-part is evaluated. In order to execute the first step, a short processing sequence of operations is applied on the currently considered fragment of text, including for example the protasis or the apodosis.
2. Whenever the if-part is satisfied, the if-category has been successfully identified.

It has to be noticed that the conditions of the if-part of the detection rules we have identified are mutually exclusive, i.e., only one rule may be satisfied on the same fragment of text, therefore no conflict may arise. As a result, the identification of if-category is univocal.

3.4 Rule 1 for *if-clauses* identification

The procedure of operations used to evaluate the if-part of the rule for detecting uncertainty in the first subcategory (*if-clauses* in a narrow sense, see Sect. 2.1) includes the following steps:

1. Checking the presence of "if" tokens.
2. Identifying protasis and apodosis.
3. Identifying the verb tenses and moods in the protasis and apodosis.
4. Check if one of the four conditions described in Table 2 applies.

If step 4 is satisfied, the corresponding type of conditional has been detected.

3.5 Rule 2 for *if-less clauses* identification

In order to detect the second subcategory (*if-less* clauses), the following operations are used:

1. Identifying all sentences with subject-verb inversion.
2. Discarding all sentences with subject-verb inversion that contain a question mark.

Table 3 Fourth subcategory rules

Case	Preceding verb tense *if/whether*	Following verb tense *if/whether*
1	Present	Present
2	Present	Past
3	Past	Present
4	Present/Past/Future	Future

3. Detecting all sentences resulting from step 2 (i.e., those sentences with subject-verb inversion that do not contain a question mark) that have as their main verb either *should* (present conditional) or *were* (simple past) or *had* (past perfect) (see *if-less* clauses, Sect. 2.1).

The verbs identified in step 3 are then tagged as UMs.

3.6 Rule 3 for *as-if/as-though* identification

In order to detect the third subcategory (*as-if* and *as-though*) the following processing step is used (see Sect. 2.1):

1. Detecting all sentences that have an *as* token immediately before an *if* or *though* token.

The pair of tokens *as-if* or *as-though* identified are tagged as UMs.

3.7 Rule 4 for *if/whether* identification

In order to detect the fourth subcategory (*if* and *whether* introducing embedded questions) the following operations are executed:

1. Checking the presence of *if/whether* tokens.
2. Identify the verbs tenses preceding and following *if/whether* tokens.
3. Checking if the verbs tenses preceding and following the *if/whether* tokens are those described in the four cases included in the following Table 3.

Rules 1 and 2 take into account also the inversion of the clauses preceding and following the *if* or *whether* token (see Sect. 2.1).

3.8 Experiments and results

The following results have been obtained by comparing the manually assigned tags with the automatically assigned ones present in the whole corpus of 80 articles. The

Table 4 If-category detection results

Precision	Recall	F1
82.64%	83.80%	83.22%

Table 5 Detection results for each if-subcategory

	Precision	Recall	F1
If-clauses in a narrow sense	82.24%	90.26%	86.06%
If-less clauses	89.47%	70.83%	79.06%
As if and as though	83.84%	100%	90.90%
If and whether	85.53%	74.71%	79.75%

comparison is performed by the Statistical Analysis subsystem (see Sect. 3.2) which also provides precision, recall, and F_1 scores.

Table 4 shows the global results for the whole *if*-category detection process.

These results significantly improve our previous ones, where precision was 26.42, recall 93.99, and F1 score 41.24.

Tables 5 shows the results for each *if*-subcategory. Such results cannot be compared with our previous study, where the if-category was not split into the subcategories described in Sect. 3.2.

The results show that our approach has high precision, recall, and F1 score in the case of *if-clauses in a narrow sense*.

In the case of *if-less clauses* we reach high precision, good F1 score, and quite good recall.

For the *as if* and *as though* case, the results feature high precision and F1 score, as well as excellent recall. In the BMJ corpus only 8 occurrences of *as if/as though* are present; the algorithm was able to detect all of them correctly, including also one occurrence in which the *as* and the *if* do not form a grammatical and semantic unit:

> (20) *"It is a practical point of some importance that no excess of anaesthetic fluid should be left immediately under the skin, as if it is [left immediately under the skin] there is a probability that the Saugmann's needle may be blocked and the manometric oscillations consequently interfered with"* (Lucas 1915).

In this excerpt, the *as* and the *if* do not form a grammatical and semantic unit, since the former has a causal meaning ('since', 'because', etc.) and the latter introduces the protasis (*if it is [left immediately under the skin]*) of an *if-clause* whose apodosis is *there is a probability that*… In other terms, in this example an *if-clause* (protasis + apodosis) is within a causal proposition introduced by *as*. In order to make the sentence unambiguous, it would have been better if the writer had inserted a comma between the *as* and the *if* and another one after the *is* (= as, if it is, there is a probability…).

Finally, the results in Table 5 show that we are able to detect the *if* and *whether* case with high precision, good recall and F1 score.

4 Conclusion, discussion and future work

Although much research has been carried out on uncertainty markers detection in the biomedical field by the NLP community, as far as we know, no study has been conducted specifically on a morphosyntactic structure, such as the if-category. As a matter of fact, while the present work focuses exclusively on the if-category detection, which includes four different sub-categories, other recent studies dealing with uncertainty, such as Jean et al. (2016), Adel and Schütze (2017), and Chen et al. (2018), show their overall results without distinguishing between the scores obtained for each category of uncertainty markers that they take into consideration. Actually, only in Jean et al. (2016) *if* and *conditionals* are taken into account, but also in their work no specific score for such markers is presented.

For this reason, it is impossible to compare both the results and the performance of the methods presented by the above-mentioned works for detecting the if-category with that presented in our work.

The approach to the automatic annotation of the if-category described in the present article performs better than the one used in our previous study (Bongelli et al. 2012) and significantly improves those results. Specifically,

- ***precision*** is always higher than 80% in all subcategories; in particular, for the *if-less* subcategory it reaches 90%.
- ***recall*** ranges from a minimum of 70.83% (for the *if-less* subcategory) to a maximum of 100% (for the *as if/as though* subcategory).
- ***F1*** score is around 80% in all four subcategories, with a minimum of 79.06% for the *if-less* subcategory and a maximum of 90.90% for the *as if/as though* subcategory.

The global scores displayed in Tables 4 and 5 show that the algorithm is able to detect all the if-subcategories with high level of precision, recall, and F1.

The main reason for false positive and false negative annotations is mainly due to the sentence complexity: when sentences are complex, the identification of the verbs involved in the if-clause, if/whether and if-less sentences can be a difficult task.

An example for a complex sentence with a false negative if-clause annotation is the following:

> (21) "If adhesion does not exist, however, there should, with proper precautions, be very little risk of infecting the surrounding peritoneal surface." (Barling 1893).

In the above example, the algorithm does not identify *should be* as a conditional modal verb, i.e., as a unit, because of the inserted adverb *however* before *there should* and of the inserted noun phrase *with proper precautions* between *should* and

be. Since the verb of the protasis is a simple present and that of the apodosis is a present conditional (= mixed type conditional), the if-clause should be detected.

An example for a complex sentence with a false positive if-clause annotation is the following:

> (22) "If, as has been repeatedly alleged, cocculus Indicus is extensively employed by dishonest brewers to impart bitterness and inebriating qualities to the pernicious liquids which they sell as beer, it is certainly eminently desirable that we should accurately ascertain the effects upon health of such a dietetic counterfeit" (Lister 1880).

In the above example, the algorithm identifies erroneously the verbal expression "s*hould accurately ascertain*…" as the apodosis verb of the if-clause sentence instead of "*it is certainly eminently desirable*…". Since the verbs of the protasis and apodosis are both in the simple present (= zero conditional), the if-clause should not be detected.

The results shown in Tables 4 and 5 disclose new perspectives to the problem of automatic detection of uncertainty markers. Considering the obtained results, we can confirm the claim introduced at the beginning of Sect. 3. Consequently, we believe that the overall process of uncertainty detection can greatly profit from a hybrid approach (see Thompson et al. 2011 and Zerva et al. 2017) which should combine:

- supervised Machine learning techniques for the most basic cases, with
- a knowledge-based approach constituted by a rule-based inference engine devoted to the if-category case and designed on the basis of the linguistic framework presented in Sect. 2.1.

In other words, while the SVM approach can reach good levels of performance for single lexical uncertainty markers identification (such as *may*, *probably*, etc.…), that are also the most frequent in our training set, a rule-based approach seems to better perform for the more complex uncertainty markers, such as the subcategories of the if-category, since the decision (classification) process is based on a sophisticated linguistic model. The rule based approach works on the sentence syntax and it is not domain-dependent; it can reach good performances also when the analysed category, as in the case of the if-category, has a little number of examples (313 occurrences in our corpus), and involves complex tasks such as (i) apodosis and protasis identification, (ii) verbs mood and tense identification.

Following this idea, for the future we plan to integrate the approach proposed in this article with that presented in Bongelli et al. 2012.

Another possible direction of investigation is focused on more sophisticated Machine Learning techniques, such as Deep Learning, like for example Recurrent Neural Networks (RNN), as described in Adel and Schütze (2017), and Chen et al. 2018, and Bidirectional Long Short-Term Memory (Bi-LSTM). The feasibility of this goal is supported by the results of other works of ours which have successfully exploited such techniques for quite sophisticated NLP tasks, such as evaluation of anaphoric references and coreference resolution (Helmy et al. 2018), and analysis of

Arabic texts for key-phrase extraction (Basaldella et al. 2016). We think that the same techniques could be successfully applied also in detecting uncertainty markers, including the if-categories.

Acknowledgements This study is a part of a research funded as PRIN (Project with National Relevance and Interest) by the Ministry of the University and Research (MIUR), Italy, named Certainty and Uncertainty in Biomedical Scientific Communication (prot. 2012C8BJ3X), coordinated by Andrzej Zuczkowski (University of Macerata) and Ramona Bongelli (University of Macerata).

References

Adel, H., & Schütze, H. (2017). Exploring different dimensions of attention for uncertainty detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics:* Volume 1, 22–34.

Agarwal, S., & Yu, H. (2010). Detecting hedge cues and their scope in biomedical literature with conditional random fields. *Journal of Biomedical Informatics, 43*(6), 953–961.

Basaldella, M., Chiaradia, G., & Tasso, C. (2016). Evaluating anaphora and coreference resolution to improve automatic keyphrase extraction, In N. Calzolari, Y. Matsumoto, and R. Prasad (Eds.), Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, (pp. 804-814), December 2016, Osaka, Japan. Publisher: The COLING 2016 Organizing Committee.

Bongelli, R., Canestrari, C., Riccioni, I., Zuczkowski, A., Buldorini, C.,Pietrobon, R., Lavelli, A., & Magnini, B. (2012) A Corpus of Scientific Biomedical Texts Spanning over 168 years annotated for Uncertainty. In Nicoletta Calzolari and Khalid Choukri and Thierry Declerck and Mehmet Uğur Doğan and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), (pp. 2009-2014). http://www.lrec-conf.org/proceedings/lrec2012/index.html.

Bongelli, R., Riccioni, I., Canestrari, C., Pietrobon, R., & Zuczkowski, A. (2014). BioUncertainty: a historical corpus evaluating uncertainty language over a 167 year span of biomedical scientific articles. In Andrzej Zuczkowski, Ramona Bongelli, Ilaria Riccioni, & Carla Canestrari (Eds.), *Communicating Certainty and Uncertainty in Medical, Supportive and Scientific Contexts* (pp. 309–339). Amsterdam/Philadelphia: Benjamins.

Bongelli, R., Riccioni, I., Burro, R., & Zuczkowski, A. (2019). Writers' uncertainty in scientific and popular biomedical articles. A comparative analysis of the British Medical Journal and Discover Magazine. PLoS ONE 14(9): 1-26. e0221933. https://doi.org/10.1371/journal. pone.0221933.

Caffi, C. (2007). *Mitigation, Studies in Pragmatics*. Amsterdam: Elsevier.

Chafe, W., & Nichols, J. (Eds.). (1986). *Evidentiality. The Linguistic Coding of Epistemology*. Norwood: Ablex.

Chen, C., Song, M., & Heo, G. E. (2018). A scalable and adaptive method for finding semantically equivalent cue words of uncertainty. *Journal of Informetrics, 12*(1), 158–180.

Crompton, P. (1997). Hedging in Academic Writing: some Theoretical Problems. *English for Specific Purposes, 16*(4), 271–287.

Dendale, P., & Tasmowski, L. (2001). Introduction. Evidentiality and related notions. *Journal of Pragmatics, 33*(3), 349–357.

Farkas, R., Vincze, V., Móra, G., Csirik, J., & Szarvas, G. (2010). The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, Uppsala, Sweden (pp. 1-12).

Fraser, B. (1980). Conversational mitigation. *Journal of pragmatics, 4*(4), 341–350.

Helmy M., Vigneshram R.M., Serra G., & Tasso C. (2018). Applying Deep Learning for Arabic Keyphrase Extraction". Procedia Computer Science 2018, 142: 254–261. Proceedings of the 4th International Conference on Arabic Computational Linguistics (ACLing 2018), November 17-19 2018, Dubai, UAE.Heritage, J. (2012). Epistemics in Action: Action Formation and Territories of Knowledge. Research on Language and Social Interaction, 45(1), 1-29.

Holmes, J. (1984). Modifying illocutionary force. *Journal of pragmatics, 8*(3), 345–365.

Hyland, K. (1994). Hedging in Academic Writing and EAP Textbooks. *English for Specific Purposes, 13*(3), 239–256.

Hyland, K. (1995). The Author in the Text: hedging Scientific Writing. *Hong Kong Papers in Linguistics and Language Teaching, 18,* 33–42.

Hyland, K. (1998a). *Hedging in Scientific Research Articles*. Amsterdam/Philadelphia: John Benjamins Publishing.

Hyland, K. (1998b). Boosting, hedging and the negotiation of academic knowledge. *Text, 18,* 349–382.

Jean, P. A., Harispe, S., Ranwez, S., Bellot, P., & Montmain, J. (2016). Uncertainty detection in natural language: A probabilistic model. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics* (p. 10). ACM.

Kärkkäinen, E. (2003). *Epistemic stance in english conversation: a description of its interactional functions, with a focus on "I Think"*. Amsterdam/Philadelphia: John Benjamins Publishing.

Kim, J.D., Ohta, T., Pyysalo, S., Kano, Y., & Tsujii, J.I. (2009). Overview of BioNLP'09 Shared Task on Event Extraction. In Proceedings of the BioNLP 2009 Workshop, Boulder, Colorado (pp. 1-9).

Kilicoglu, H., & Bergler, S. (2008). Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics, 9*(11), S10.

Kilicoglu, H., & Bergler, S. (2010). A High-Precision Approach to Detecting Hedges and Their Scopes. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, Uppsala, Sweden (pp. 70-77).

Lakoff, G. (1973). Hedges: a study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic, 2*(4), 458–508.

Lyons, J. (1977). *Semantics*. Cambridge: Cambridge University Press.

Malhotra, A., Younesi, E., Gurulingappa, H., & Hofmann-Apitius, M. (2013). 'HypothesisFinder:' a strategy for the detection of speculative statements in scientific text. *PLoS Computational Biology, 9*(7), e1003117.

Narayanan, R., Liu, B., & Choudhary, A. (2009). Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*: Volume 1 (pp. 180-189). Association for Computational Linguistics.

Nuyts, J. (2001). Subjectivity as an evidential dimension in epistemic modal expressions. *Journal of Pragmatics, 33*(3), 383–400.

Ochs, E. (1996). Linguistic resources for socializing humanity. In J. Gumperz & S. Levinson (Eds.), *Rethinking Linguistic Relativity* (pp. 407–437). New York: Cambridge University Press.

Özgür, A., & Radev, D.R. (2009). Detecting speculations and their scopes in scientific text. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*: Volume 3 (pp. 1398-1407). Association for Computational Linguistics.

Palmer, F. (1986). *Mood and modality*. Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Radden, G., & Dirven, R. (2007). *Cognitive English Grammar*. Amsterdam/Philadelphia: John Benjamins.

Rubin, V.L. (2007). Stating with Certainty or Stating with Doubt: Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements. *Proceedings of NAACL HLT 2007*, Companion Volume, 141–144.

Salager-Meyer, F. (1994). Hedges and textual communicative function in medical english written discourse. *English for Specific Purposes, 13*(2), 149–170.
Stivers, T., Mondada, L., & Steensig, J. (2011). *The Morality of Knowledge in Conversation*. Cambridge: Cambridge University Press.
Szarvas, G., Vincze, V., Farkas, R., Móra, G., & Gurevych, I. (2012). Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistic., 38*(2), 335367.
Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2011). Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics, 12*(393), 1–18.
Vaihinger, H. (1952). The philosophy of 'as if': *A System of the Theoretical, Practical and Religious Fictions of Mankind*. London: Routledge & Kegan Paul. (Original work published 1911, Die Philosophie des Als Ob).
Velldal E., Øvrelid, L., & Oepen, S. (2010). Resolving Speculation: MaxEnt Cue Classification and Dependency-Based Scope Rules. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task,* Uppsala, Sweden (pp. 48-55).
Velldal, E., Øvrelid, L., Read, J., & Oepen, S. (2012). Speculation and negation: rules, rankers, and the role of syntax. *Computational Linguistics, 38*(2), 369–410.
Vincze, V., Szarvas, G., Farkas, R., Móra, G., & Csirik, J. (2008). The bio-scope corpus: biomedical texts annotated for uncertainty negation and their scopes. *BMC Bioinformatics, 9*(11), S9.
Willett, T. (1988). A cross-linguistic survey of the grammaticalization of evidentiality. *Studies in Language, 12*(1), 51–97.
Zerva, C., Batista-Navarro, R., Day, P., & Ananiadou, S. (2017). Using uncertaninty to link and rank evidence from biomedical literature for model curation. *Bioinformatics, 332*(23), 3784–3792.
Zhou, H., Huang, D., Li, X., & Yang, Y. (2011). Combining structured and flat features by a composite Kernel to detect hedges scope in biological texts. *Chinese Journal of Electronics, 20*(3), 476–482.
Zhou, H., Deng, H., Huang, D., & Zhu, M. (2015). Hedge scope detection in biomedical texts: an effective dependency-based method. *PLoS ONE, 10*(7), 1–16.
Zou, B., Zhou, G., & Zhu, Q. (2013). Tree Kernel-based Negation and Speculation Scope Detection with Structured Syntactic Parse Features. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 968-976).
Zuczkowski, A., Colella, G., Riccioni, I., Bongelli, R., & Canestrari, C. (2014a). Italian come se "as if": evidential and epistemic aspects. In Sibilla Cantarini, Werner Abraham, & Elisabeth Leiss (Eds.), *Certainty-uncertainty—and the attitudinal space in between* (pp. 297–323). Benjamins: Amsterdam/Philadelphia.
Zuczkowski, A., Bongelli, R., Vincze, L., & Riccioni, I. (2014b). Epistemic stance: knowing, unknowing, believing (KUB) positions. In Andrzej Zuczkowski, Ramona Bongelli, Ilaria Riccioni, & Carla Canestrari (Eds.), *Communicating certainty and uncertainty in medical, supportive and scientific contexts* (pp. 115–136). Benjamins: Amsterdam/Philadelphia.
Zuczkowski, A., Bongelli, R., Riccioni, I., Valotto, M., Burro, R. (2016). Writers' uncertainty in a corpus of scientific biomedical articles with a diachronic perspective. In esús Romero-Trillo (Ed.), Yearbook of Corpus Linguistics and Pragmatics 2016.Global Implications for Society and Education in the Networked Age (pp. 203-241). Springer International Publishing.
Zuczkowski, A., Bongelli, R., & Riccioni, I. (2017). *Epistemic stance in dialogue*. Benjamins: Amsterdam/Philadelphia.

BMJ References

Barling, G. (1893). Appendicitis: an analysis of sixty-eight cases, with comments and a summary of the conditions requiring operation. *BMJ, 1*(1686), 838–841.
Barr, J. (1909). An address on the treatment of chronic degenerative lesions of the heart and aorta. *BMJ, 2*(2532), 61–64.
Bourne, G. (1955). Effects of flying on patients with cardiovascular disease. *BMJ, 1*(4909), 310–313.
Burn, J. H., & Rand, M. J. (1958). Action of nicotine on the heart. *BMJ, 1*(5063), 137–139.
Crichton-Browne, J. (1875). On the actions of picrotoxine, and the antagonism between picrotoxine and chloral hydrate. *BMJ, 1,* 540–542.
Dally, J. H. (1913). A clinical lecture on maximal and minimal blood pressures and their significance. *BMJ, 2*(2754), 899–901.

Davies, W. (1913). A method of operating for radical cure of inguinal hernia. *BMJ, 2*(2751), 727–728.
Doll, R., & Hill, A. B. (1956). Lung cancer and other causes of death in relation to smoking; a second report on the mortality of British doctors. *BMJ, 2*(5001), 1071–1081.
Horton, R. E. (1956). Use of grafts in treatment of atherosclerosis of lower limbs. *BMJ, 1*(4958), 81–82.
Johnson, G. (1865). Notes on the pathology and treatment of cholera. *BMJ, 2*(253), 465–466.
Lister, J. (1880). Remarks on micro-organism: their relation to disease. *BMJ, 2*(1027), 363–365.
Lowther, C., & Turner, R. (1963). Guanethidine in the treatment of hypertension. *BMJ, 2*(5360), 776–781.
Lucas, G. (1915). The treatment of pulmonary tuberculosis by nitrogen compression. *BMJ, 2*(2849), 211–213.
Radford, T. (1849). A successful case of caesarean section, with remarks. *BMJ, 13,* 456–460.
Scott-Coombes, D., & Thompson, J. (1991). Bile duct stones and laparoscopic cholecystectomy. *BMJ, 303*(6813), 1281–1282.
Wells, S. (1861). Lecture on vesico-vaginal and rectovaginal fistula. *BMJ, 2*(35), 223–225.
Wilson, G. S. (1947). The value of vaccination in control of tuberculosis. *BMJ, 2*(4534), 855–859.