A Corpus of Clinical Practice Guidelines Annotated with the Importance of Recommendations

Jonathon Read[♠] **Erik Velldal**[♡] **Marc Cavazza**[♢] **Gersende Georg**[♣]

[♠] Teesside University, School of Computing

[♡] University of Oslo, Department of Informatics

[♢] University of Kent, School of Engineering and Digital Arts

[♣] Haute Autorité de Santé

j.read@tees.ac.uk, erikve@ifi.uio.no, m.o.cavazza@kent.ac.uk, g.georg@has-sante.fr

Abstract

In this paper we present the Corpus of REcommendation STrength (CREST), a collection of HTML-formatted clinical guidelines annotated with the location of recommendations. Recommendations are labelled with an author-provided indicator of their strength of importance. As data was drawn from many disparate authors, we define a unified scheme of importance labels, and provide a mapping for each guideline.

We demonstrate the utility of the corpus and its annotations in some initial measurements investigating the type of language constructions associated with strong and weak recommendations, and experiments into promising features for recommendation classification, both with respect to strong and weak labels, and to all labels of the unified scheme. An error analysis indicates that, while there is a strong relationship between lexical choices and strength labels, there can be substantial variance in the choices made by different authors.

Keywords: Corpus annotation, Extra-propositional aspects of meaning, Normalised Pointwise Mutual Information, Support Vector Machines

1. Introduction

Practice guidelines document various recommendations regarding best practice for particular clinical situations. They provide assistance to medical practitioners and their patients when making decisions about individual health care (Field and Lohr, 1990), and can also influence high-level health care policy making.

Guideline authors often annotate their recommendations with labels that indicate the strength of importance of the statement (see Figure 1 for examples). Such labels present an interesting resource that can add a further dimension to current work in extra-propositional aspects of meaning (Morante and Sporleder, 2012; Blanco et al., 2015)—namely the investigation of linguistic constructions that convey the importance of propositions. To support such work we have developed the Corpus of REcommendation STrength (CREST), a collection of clinical guidelines annotated with instances of recommendations together with their strength of importance as specified by their authors.[1]

In this paper we introduce related work in Section 2, before describing how we created our corpus in Section 3. Section 4 presents an initial analysis of some of the features of the strength of importance. Then in Section 5 we present the results and an error analysis of some initial classification experiments including the discrimination between the two most frequently occurring labels as they multi-class problem involving all labels. We then conclude in Section 6 and outline our plans for further development and investigation of the corpus.

[1]The corpus is freely available to download from `http://www.velldal.net/erik/crest/`

2. Related work

Zhang and Xu (2015) created a corpus of emails annotated with their importance, where participants solicited through Mechanical Turk (rather than the authors) chose labels of important, normal and not important. The authors found that a machine learning-based classifier could identify unimportant emails with a precision of 54.7%.

Hussain, Michel, and Shiffman (2009) collected a representative sample of guideline recommendations with the aim of understanding how recommendations are written. They noted inconsistencies in lexical and typographical indicators, such that almost a third of guidelines contained recommendations that were not clearly identifiable. This corpus differs from our own in that: it contains only a subset of the recommendations in each guideline considered whereas ours contains all recommendations; it retains recommendation strength as in-text annotations rather than the machine-readable versions in our corpus; and it (to our knowledge) is not freely-available.

Lomotan, Michel, Lin, and Shiffman (2009) investigated how members of the health services community interpret the strength of deontic expressions in clinical guidelines. Participants in this study judged the level of obligation conveyed by twelve deontic expressions situated in otherwise identical sentences, rating obligation from 0 (no obligation) to 100 (full obligation). Participants interpreted the deontic expression "must" as conveying the highest level of obligation, while while "may" and "may consider" indicated the lowest. Lomotan et al. offer an interesting evaluation of health professionals' interpretation of deontic expressions; the corpus presented in this paper can facilitate an extension

of this research by considering the interaction of deontic expressions with other extra-propositional aspects of meaning such as speculation (Velldal et al., 2012), and by analysing associations of linguistic features with the explicit labels of recommendation strength.

3. The corpus

The initial CREST release contains 170 guidelines developed by 69 different institutions. These institutions employ 30 different schemes for the rating of recommendations, with variations in grade labels and descriptions. Below we discuss a unified rating scheme we use for categorising recommendations in the corpus (Section 3.2), after discussing details about the extraction process (Section 3.1).

3.1. Data collection

The guidelines were obtained from the National Guideline Clearinghouse,[2] a public database maintained by the Agency for Healthcare Research and Quality, of the U.S. Department of Health and Human Services. Using the web interface to search for guidelines where the strength of recommendations are weighted according to a rating scheme yielded 1,808 guidelines; for this initial build of the corpus we acquired the first 180 guidelines. Of these 10 were discarded because they did not explicitly label recommendations with weights (contrary to their metadata). The rating scheme and section containing the recommendations were extracted from the remaining guidelines, before manually replacing each textual label of recommendation strength with an attribute attached to the HTML element that corresponded to the scope of the annotation (most typically `<p/>` or `<li/>`). In some cases labels of importance scope subsententially; this was represented by introducing `<span/>` elements to record the scope of labels. Finally, the in-text annotations of strength were removed, along with evidence-level[3] annotations, if present (under the assumption that evidence levels could correlate with recommendation strengths and thus become informative but extra-linguistic features of importance). Figure 1 presents a typical example of the manual transformation process.

To promote comparability in future research, the corpus includes suggested partitions, wherein 30% of guidelines are reserved for held-out testing and the remainder are allocated into ten folds for cross-validation in developmental experiments.

3.2. Unified rating scheme

The 170 recommendation guidelines in the corpus employ 30 different rating schemes denoting various subtleties about recommendation strength. The schemes take on a variety of labels and levels of granularity (e.g. some consider only strong and weak recommendations, while others include three or four levels). For instance, the GRADE system (Guyatt et al., 2008) distinguishes between strong recommendations (benefits of action clearly outweigh its risks/burdens) and weak recommendations (the balance between benefits and risks/burdens is less clear). Other guideline schemes offer finer granularity between grades of recommendation strength. Some schemes also single out special recommendations that are based not on scientific evidence but instead on the consensus of experts. Other schemes include a label to indicate that a recommendation could not be made at all because of inconclusive evidence. For instance, the U. S. Preventive Task Force grades strength with the letters A–C, with D to indicate recommendations against an action, and I to highlight that evidence is insufficient to make any recommendation for the given question (U.S. Preventive Services Task Force, 2014). Strength grades are often expressed in terms of the balance between expected benefit and harm for those patients to which the recommendation applies, but some instead are defined according to the quality of evidence upon which the recommendation is based.

Such disparity in strength schemes presents a challenge for research into the choice of linguistic constructions expressing recommendation strength because differences in naming conventions and granularity make it difficult to apply machine learners for making generalisations across schemes. To alleviate these challenges, we map the disparate schemes into one unified scheme:

- The labels STRONG, MODERATE, and WEAK are assigned based on the strength of the associate recommendation (strong recommendations being the most important while weak recommendations are least important).
- The CONSENSUS consensus label is assigned because, while there is insufficient evidence to label a recommendation as strong, moderate, or weak, there is nevertheless agreement among a committee of experts that the associated action is appropriate.
- The INCONCLUSIVE label is assigned because there is insufficient evidence to label a recomemdnation as strong, moderate or weak, and the committee of experts was unable to reach a consensus on what is an appropriate action.

Figure 2 enumerates the labels and associated descriptions for two typical schemes for recommendation strength and the unified labels into which they map.

Mappings were assigned by looking at the grades of a scheme in relation to each other, together with their descriptions, and considering the following heuristics in parallel:

- If a unified label matches or is synonymous with the label of a grade or keywords in its description it is a good candidate for mapping.
- Strength can be inferred from a position of a label relative to other labels in the scheme. For example, a scheme using A, B and C suggests the labels might be respectively interpreted as strong, moderate and weak.

[2] `http://www.guideline.gov/`

[3] Guidelines are often developed through the evidence-based medicine framework for systematically assessing clinical research findings; some guidelines supplement the strength annotation with an indication of the level of quality of evidence upon which the recommendation is based.

```
<p>Clinicians might counsel patients
that this symptomatic benefit is
possibly maintained for 1 year (Level
C), although THC is probably ineffective
for improving objective spasticity
measures (short-term) or tremor (Level
B).</p> <p>Clinicians might offer
Sativex oromucosal cannabinoid spray
(nabiximols), where available, to reduce
symptoms of spasticity, pain, or urinary
frequency, although it is probably
ineffective for improving objective
spasticity measures or number of urinary
incontinence episodes (Level B).</p>
```

```
<p>Clinicians <span recommendation="C">
might counsel patients that this
symptomatic benefit is possibly
maintained for 1 year</span>, <span
recommendation="B">although THC is
probably ineffective for improving
objective spasticity measures
(short-term) or tremor</span>.</p> <p
recommendation="B">Clinicians might
offer Sativex oromucosal cannabinoid
spray (nabiximols), where available,
to reduce symptoms of spasticity, pain,
or urinary frequency, although it is
probably ineffective for improving
objective spasticity measures or number
of urinary incontinence episodes.</p>
```

Figure 1: An example demonstrating the manual annotation process. Primary HTML is shown on the left, in which recommendation strength is annotated textually. The resulting XML is on the right, where recommendation strength is recorded using attributes on elements, with `<span/>` elements introduced to indicate subsentential scoping where necessary.

(a)

Label	Description	Mapping
A	There is good evidence to recommend the clinical preventive action.	STRONG
B	There is fair evidence to recommend the clinical preventive action.	WEAK
C	The existing evidence is conflicting and does not allow a recommendation for or against use of the clinical preventive action; however other factors may influence decision-making.	INCONCLUSIVE
D	There is fair evidence to recommend against the clinical preventive action.	WEAK
E	There is good evidence to recommend against the clinical preventive action.	STRONG
L	There is insufficient evidence (in quantity or quality) to make a recommendation; however, other factors may influence decision-making.	INCONCLUSIVE

(b)

Label	Description	Mapping
A	The USPSTF recommends the service. There is a high certainty that the net benefit is substantial.	STRONG
B	The USPSTF recommends the service. There is a high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial.	MODERATE
C	The USPSTF recommends selectively offering or providing this service based on professional judgment and patient preferences. There is at least moderate certainty that the net benefit is small.	WEAK
D	The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits.	MODERATE
I	The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, or of poor quality, or conflicting, and the balance of benefits and harms cannot be measured.	INCONCLUSIVE
EO	Expert opinion.	CONSENSUS

Figure 2: Grades in two example recommendation strength schemes, with mappings into our unified scheme.

Label	Instances	Length	Types
STRONG	1,756	33.9 ± 26.8	5,219
MODERATE	373	38.6 ± 26.9	2,378
WEAK	1,358	36.0 ± 34.5	4,926
CONSENSUS	364	31.4 ± 23.1	1,905
INCONCLUSIVE	178	33.7 ± 20.3	1,300
overall	*4,029*	*34.8 ± 29.2*	*8,138*

Table 1: The composition of the corpus, in terms of the number of instances of each strength of recommendation in the unified scheme, their average length (with standard deviation indicated using ±), and vocabulary size.

This is not necessarily the case however (such as in Figure 2a where the descriptions for labels C, D, E and L overrule this heuristic).

- The polarity of a recommendation description should not affect its unified label (such as for labels D and E in Figure 2a).

The mappings are non-destructive as they are specified as a series of unified equivalents associated with each grade definition.

To assess the reliability of such mappings we conducted an inter-annotator agreement study where participants considered scheme labels and their descriptions and selected the most appropriate grade in the unified strength scheme. Using Fleiss' (1971) Kappa, we found strong agreement between untrained annotators ($\kappa = 0.77$, $n = 14$), suggesting that a unified scheme is sufficiently reliable, at least in terms of interpretation by respondents who are not experts in the clinical domain. Interestingly, measuring agreement between two authors of this paper indicated that experience with linguistic analysis does not necessarily result in significantly greater agreement as evidenced by Cohen's (1960) Kappa ($\kappa = 0.80$). Future work will expand this study to consider agreement between health care professionals.

The labels of our unified scheme are listed in Table 1, along with various statistics describing the distribution of recommendations in the corpus.

4. Features of recommendation strength

Some features of recommendation strength may be enumerated through introspection; most readers of English can discern the difference in the use of *must*, *should*, or *may*, for example. This corpus presents an opportunity to investigate linguistic constructions that are less obviously associated with the strength of a recommendation. To initiate investigations in this direction we estimated the association between various linguistic features (f) and the labels of the unified strength scheme (l) using normalised pointwise mutual information (Bouma, 2009):

$$\text{NPMI}(f;l) = \log \frac{p(f,l)}{p(f)\,p(l)} \times \frac{1}{-\log p(f,l)}$$

NPMI falls in the range $[-1, 1]$ where 1 indicates perfect positive correlation, -1 indicates perfect negative correlation, and 0 indicates independence. In our experiments we computed probability using maximum likelihood estimation based on the number of recommendations containing a feature, as observed in the development corpus.

Adopting the semantic orientation approach introduced by Turney (2002) for sentiment polarity, we estimate the strength conveyed by a feature by finding the difference[4] between its association with STRONG and WEAK labels:

$$\text{strength}(f) = \frac{1}{2}\left(\text{NPMI}(f;\text{STRONG}) - \text{NPMI}(f;\text{WEAK})\right)$$

We used the strength measure to score each of the twelve expressions investigated in Lomotan, Michel, Lin, and Shiffman's (2009) survey of health professionals' interpretation of deontics. The resulting scores correlated reasonably strongly with the 'level of obligation' scores found by Lomotan et al. (Pearson's $r = 0.56$), indicating that there are similarities between the results of our respective methodologies.

Table 2 lists the twenty highest scoring trigrams found using this measure, where trigrams are formed from lemmas predicted by Stanford's CoreNLP (Manning et al., 2014). The trigram *be reasonable in* suggests a deontic expression not covered by the survey of Lomotan et al. (2009), as in "action X is reasonable for all patients". Other features are not deontic expressions yet seem reasonable indicators of recommendation strength—for example, *all patient with* narrows the focus of a recommendation to all patients of a particular class. Some trigrams seem less intuitive cues of recommendation strength and are perhaps indicative of extra-linguistic phenomena (e.g. *[S] the uspstf* having a high postive strength suggests that USPSTF guideline developers are more inclined to offer STRONG recommendations).

5. Initial classification experiments

To explore the difficulty of automatically predicting the strength labels of recommendations we performed some preliminary classification experiments. Section 5.1 reports results of classification using the two most frequent classes (STRONG and WEAK) together with an error analysis. Section 5.2 investigates the multi-class problem with experiments involving all five labels.

5.1. Discriminating between STRONG and WEAK recommendations

Using Joachims' (1999) implementation of support vector machines[5] we attempted to discriminate between recommendations that were either STRONG or WEAK (according to the unified scheme discussed in Section 3.2). We evaluated the classification accuracy using ten-fold cross validation on guidelines from the development set, and describing

[4] If, when estimating feature strength, there are no instances of a feature for a given label observable in the corpus we assume they occur independently (i.e. $\text{NPMI}(f;l) = 0$).

[5] SVM^{light} — available from `http://svmlight.joachims.org/`

Trigram	Frequency	Strength
write group recommend	128	0.640
guideline panel suggest	37	-0.601
may be consider	38	-0.591
it be suggest	27	-0.543
[S] consider use	21	-0.497
[S] clinician may	25	-0.496
[S] there be	44	-0.397
all patient with	29	0.369
be recommend for	49	0.365
be indicate in	24	0.324
be recommend .	43	0.323
there be insufficient	36	-0.287
it be recommend	45	0.286
be recommend in	39	0.285
be recommend that	43	0.276
aged $\geq$ [NUM]	31	0.272
be reasonable in	27	0.272
should not be	54	0.259
[S] the uspstf	24	0.257
in the individual	20	-0.254

Table 2: Top twenty trigrams of lemmas occurring in the corpus, ordered by absolute strength of importance as calculated using the strength function. [S] represents the start of a sentence; [NUM] is wildcard where any number applies.

recommendations using features of word forms, lemmas, lemmas with part-of-speech, bigrams, trigrams, and dependencies predicted using Stanford's CoreNLP (Manning et al., 2014), and the trade-off between training error and margin was set automatically by the machine learning package. Table 3 lists the results of the support vector machines compared with two baseline approaches to classification:

Majority always selecting the more frequent class, STRONG (57.3%); and

Majority & Deontic Verbs using the properties of deontic verbs by selecting STRONG when observing a verb of obligation and WEAK when observing a verb of permissibility, and backing off to the majority baseline when neither is observed (61.4%),

Support vector machines trained with lemmas offer a significant improvement in classification accuracy (75.7%).

An error analysis (conducted on 20% of the errors made by support vector machines trained with lemma features) suggested that guidelines are somewhat inconsistent with respect to lexical choices made for STRONG and WEAK recommendations. The majority (38.5%) of errors analysed appeared to be due to guideline authors using deontic verbs one might normally associate with the opposite class, for example:

Features	Accuracy
Baselines	
Majority	57.3
Majority & Deontic Verbs	61.4
Support Vector Machines	
Words	74.2
Lemmas	75.7
Lemmas with Part-of-Speech	75.6
Bigrams	73.7
Trigrams	68.7
Dependencies	71.8

Table 3: The cross-validated accuracy of baselines and of support vector machines when distinguishing between STRONG and WEAK recommendations in the development subset.

(1) The guideline developers suggest that other methodologies for repackaged IVFE, such as drawn-down IVFE units, are preferable. [Annotated as STRONG, predicted as WEAK]

(2) Patients should receive a maximum trial period of 16 weeks of therapy. This should comprise 8 weeks at the starting dose of ESA G-CSF and a further 8 weeks at the higher doses, if required. [Annotated as WEAK, predicted as STRONG].

Conversely, 18.9% of errors occur in sentences that use deontic verbs that are associated with their label, and yet were misclassified, for example:

(3) TEE should be performed in patients considered for percutaneous mitral balloon commissurotomy to assess the presence or absence of left atrial thrombus and to further evaluate the severity of mitral regurgitation (MR). [Annotated as STRONG, predicted as WEAK]

(4) A suggestion is made to use ethanol lock to prevent catheter-related bloodstream infections (CLABSI) and to reduce catheter replacements in children at risk of PNALD. [Annotated as WEAK, predicted as STRONG]

These apparently contradictory types of errors point to noise in the lexical choices associated with labels assigned by the guideline authors; a challenge of this dataset is to automatically identify and handle such noisy examples.

Sentences that do not contain verbs associated with recommendation strengths also account for several errors. The imperative form occurred frequently in the error analysis (25%), for example:

(5) Assess for deterioration of the ulcer or possible infection when the individual reports increasing intensity of pain over time. [Annotated as STRONG, predicted as WEAK]

(6) Ensure that a complete skin assessment is part of the risk assessment screening policy in place in all health care settings. [Annotated as WEAK, predicted as STRONG]

The imperative form appeared more frequently in association with STRONG recommendations however—suggesting that its presence may be a useful feature in future classification experiments.

Other errors (10.8%) were concerning recommendations that communicate facts and definitions rather than actions:

(7) Several studies report changes in nutrient adequacy with caloric restriction, however the extent of nutrient inadequacy and the nutrients affected are dependent on the composition of the diet followed, as well as on the nutritional needs of the individual. [Annotated as STRONG, predicted as WEAK]

(8) Obesity (body mass index [BMI] greater than 30kg/m2) is a condition for which there is no restriction on the use of the progestogen-only implant. [Annotated as WEAK, predicted as STRONG]

Such recommendations contain few explicit cues as to the importance of the reported facts and definitions.

5.2. Discriminating between all labels

We next performed multi-class classification experiments, using Joachims' implementation of multi-class support vector machines (Crammer and Yoram, 2001).[6] Table 4 lists the cross-validated results of training using lemma features. We found a macro-averaged F_1 of 31.1% and a micro-averaged F_1 of 58.0% (which can be compared to the majority-choice baseline macro-average of 20.0% and micro-average of 44.1%).

An error analysis yielded similar findings to those reported above, with the additional observation that the infrequent classes (MODERATE, CONSENSUS, and INCONCLUSIVE) were rarely predicted by the classifier—as can be seen in the contingency table shown in Table 5, they account for only 3.2% of predictions despite occurring with 23.1% of recommendations. Use of cost-sensitive learning can yield significant gains in performance for multi-class text classification problems, especially in case of unbalanced training data such as this (Read et al., 2012).

6. Conclusions and outlook

This paper has described the development of the Corpus of REcommendation STrength (CREST), a collection of clinical guidelines annotated with recommendations and their corresponding strength of importance. The field of NLP has in recent years seen increasing interest in extra-propositional aspects of meaning. This sub-area of research includes phenomena like subjectivity, hedging or uncertainty, negation, factuality, and other related phenomena (Morante and Sporleder, 2012; Blanco et al., 2015). The release of CREST facilitates a new direction within this line of research, focusing on linguistic constructions that convey the importance of propositions.

The data set includes the full original HTML of the guidelines in addition to the sections of major recommendations extracted from each guideline mapped to a shared scale of strength grading. While the initial release of the corpus contains English guidelines only, we will expand it to also include guidelines written in French. The corpus also comes with a pre-defined training and development section (of ten folds) and a test set.

This corpus presents some interesting challenges for machine learning-based classification, and one can immediately observe that it may require more sophisticated representations than simple bags-of-words. The various phenomena filed under extra-propositional aspects of meaning interact in non-trivial ways. For example, in:

(9) Clinicians might counsel patients that this symptomatic benefit is possibly maintained for 1 year.

the hedge *possibly* affects the described benefit and not the strength of the recommendation. The full version of this paper will discuss the challenges in applying machine learning to identifying the strength of importance, building on preliminary experiments assessing the contribution of various linguistic features such as part-of-speech and dependency triples.

In the biomedical domain of the CREST data, judgments of importance play a vital role. At the same time, guideline authoring is a subjective process (Shaneyfelt and Centor, 2009); clinical practitioners may apply their own experience when interpreting evidence (Shrier et al., 2008). This can lead to inconsistency in the choice of strength la-

Strength	Precision	Recall	F_1
STRONG	57.6	83.8	68.3
MODERATE	5.7	0.8	1.4
WEAK	61.8	61.6	61.7
CONSENSUS	20.0	2.5	4.4
INCONCLUSIVE	83.3	11.2	19.8
macro-average	45.7	32.0	31.1
micro-average	58.0	58.0	58.0

Table 4: The results of predicting recommendations' labels of importance, using support vector machines trained with lemma features. Compare with a baseline of always selecting the majority class (STRONG) which achieves a macro-averaged F_1 of 20.0% and a micro-averaged F_1 of 44.1%.

[6] SVM^{multiclass} — available from https://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html

		Predicted					
		STRONG	MODERATE	WEAK	CONSENSUS	INCONCLUSIVE	*Total*
Actual	STRONG	37.0	0.4	6.4	0.3	0.1	*44.2*
	MODERATE	6.3	0.1	2.7	0.3	0.0	*9.4*
	WEAK	12.1	0.3	20.2	0.2	0.0	*32.8*
	CONSENSUS	6.1	0.6	2.3	0.2	0.0	*9.2*
	INCONCLUSIVE	2.7	0.1	1.1	0.1	0.5	*4.5*
	Total	*64.2*	*1.5*	*32.7*	*1.1*	*0.6*	

Table 5: A contingency table showing the distribution of the multi-class support vector machines classify when predicting all five labels. Rows indicate the predicted label while columns indicate the actual label. Values are percentages with respect to all recommendations.

bels, and arguably the linguistic choices made in association with the labels. Our ultimate aim in developing this corpus is to augment document engineering environments for clinical guidelines (Georg and Jaulent, 2007) with functionality that assists developers in constructing unambiguous recommendations. A traditional classification approach can help identify ambiguous recommendations, but a further requirement is that a convincing justification must be offered to the developer. For instance, the authors of the recommendation:

(10) High-flow oxygen should be administered by face mask to all patients with anaphylaxis.

assigned a strength grade described as "troublingly inconsistent or inconclusive studies"—yet, with the deontic expression *should* and its relevance specified as *all patients*, readers might take this as a strong recommendation. Detecting these sorts of inconsistencies provides just one of several use cases for experimenting with machine learning with the CREST corpus.

7. Acknowledgements

The contribution of the third and fourth authors was funded in part by the European Commission through the FP7 Open FET "MUSE" Project (ICT-296703). The contents of this paper only reflect the authors' opinions and not necessarily the official position of Haute Autorité de Santé.

8. Bibliographical References

Blanco, E., Morante, R., and Sporleder, C. (2015). *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM 2015)*.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference 2009*, pages 31–40, Tübingen, Germany.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Crammer, K. and Yoram, S. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292.

Field, M. J. and Lohr, K. N. (1990). *Clinical Practice Guidelines: Directions for a New Program*. National Academies Press.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Georg, G. and Jaulent, M.-C. (2007). A document engineering environment for clinical guidelines. In *Proceedings of the 2007 ACM Symposium on Document Engineering*, pages 69–78.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., and Schünermann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 336:924–926.

Hussain, T., Michel, G., and Shiffman, R. N. (2009). The Yale Guideline Recommendation Corpus: A representative sample of the knowledge content of guidelines. *International Journal of Medical Informatics*, 78:354–363.

Joachims, T. (1999). Making large-scale SVM learning practical. In B Schölkopf, et al., editors, *Advanced in Kernel Methods—Support Vector Learning*. MIT Press.

Lomotan, E. A., Michel, G., Lin, Z., and Shiffman, R. N. (2009). How "should" we write guideline recommendations? Interpretation of deontic terminology in clinical practice guidelines: Survey of the health services community. *Quality & Safety in Healthcare*, 19:509–513.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Morante, R. and Sporleder, C. (2012). *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM 2012)*.

Read, J., Velldal, E., and Øvrelid, L. (2012). Topic classification for suicidology. *Journal of Computing Science*

and Engineering, 6:143–150.

Shaneyfelt, T. M. and Centor, R. M. (2009). Reassessment of clinical practice guidelines: Go gently into that good night. *JAMA*, 301(8):868–869.

Shrier, I., Boivin, J.-F., Platt, R. W., Steele, R. J., Brophy, J. M., Carnevale, F., Eisenberg, M. J., Furlan, A., Kakuma, R., Macdonald, M. E., Pilote, L., and Rissignol, M. (2008). The interpretation of systematic reviews with meta-analyses: An objective or subjective process? *BMC Medical Informatics and Decision Making*, 8(1):19.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424.

U.S. Preventive Services Task Force. (2014). Grade definitions. `http://www.uspreventiveservicestaskforce.org/Page/Name/grade-definitions`, October. Accessed 2016-03-01.

Velldal, E., Øvrelid, L., Read, J., and Oepen, S. (2012). Speculation and negation: Rules, rankers and the role of syntax. *Computational Linguistics*, 38(2):368–410.

Zhang, F. and Xu, K. (2015). Annotation and classification of an email importance corpus. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 651–656.