# SUPPLEMENTAL

## DATA EXAMPLES

Some examples of sentences from the Asthma and ENT forums labeled with SC (in italics) and DT (in bold) labels are shown below:

> "Cold dry air is a common trigger , I 'm also haven't a lot of trouble keeping the *asthma* under control now that is it winter (only diganosed last spring)."
>
> "I had actually been feeling *spasms* in my throat that I thought were *palpitations* but it ended up not being my heart."
>
> "Now I have developed a low grade *fever* and *blisters* in my throat."
>
> "Would love some feedback as I 'm anxious."
>
> "No *stuffed nose* , no *discharge*."
>
> "yes i realize that i should have used **ear plugs** and yes i 've learned my lesson that i will use ***plugs*** from now on."
>
> "I have *chronic sinusitis* , *scars* on both ears from past *infections* , and "fairly severe *deviated septum,* "."
>
> "I went to the doctor and he gave me **augmittin** it cleared the *white patches* right up."
>
> "I went to the health food store and found **Wally's Ear Oil** about 2 weeks ago after reading some of the posts here."
>
> "I am interested in **Xanax** side affect of *loosing taste and smell*."
>
> "It sounds like *chronic non-infectious bronchitis*."
>
> "I've had chest x-ray - normal."
>
> "Once I had my *sinus* **surgeries** my *asthma* improved dramatically."

## STOP WORDS LIST

### Medical stop words

disease, diseases, disorder, symptom, symptoms, drug, drugs, problems, problem,prob, probs, med, meds, pill, pills, medicine, medicines, medication, medications, treatment, treatments, caps, capsules, capsule, tablet, tablets, tabs, doctor, dr, dr., doc, physician, physicians, test, tests, testing, specialist, specialists, side-effect, side-effects, pharmaceutical, pharmaceuticals, pharma, diagnosis, diagnose, diagnosed, exam, challenge, device, condition, conditions, suffer, suffering ,suffered, feel, feeling, prescription, prescribe, prescribed, over-the-counter, otc

## General stop words

a, about, above, after, again, against, all, am, an, and, any, are, aren't, as, at, be, because, been, before, being, below, between, both, but, by, can, can't, cannot, could, couldn't, did, didn't, do, does, doesn't, doing, don't, down, during, each, few, for, from, further, had, hadn't, has, hasn't, have, haven't, having, he, he'd, he'll, he's, her, here, here's, hers, herself, him, himself, his, how, how's, i, i'd, i'll, i'm, i've, if, in, into, is, isn't, it, it's, its, itself, let's, me, more, most, mustn't, my, myself, no, nor, not, of, off, on, once, only, or, other, ought, our, ours , ourselves, out, over, own, same, shan't, she, she'd, she'll, she's, should, shouldn't, so, some, such, than, that, that's, the, their, theirs, them, themselves, then, there, there's, these, they, they'd, they'll, they're, they've, this, those, through, to, too, under, until, up, very, was, wasn't, we, we'd, we'll, we're, we've, were, weren't, what, what's, when, when's, where, where's, which, while, who, who's, whom, why, why's, with, won't, would, wouldn't, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves, n't, 're, 've, 'd, 's, 'll, 'm,

# ADDITIONAL EXPERIMENTS

## Experiment 1: Labeling Data by Matching Patterns

In our paper, we learn dictionaries for SC and DT phrases and use the dictionaries to label data. The labeling is done by dictionary look-up and does not consider context. Context is only considered to learn patterns that extract new dictionary phrases. Another approach is to label data using the learned patterns, which uses only context. We compare both the approaches in Table 1. The system 'Pattern Matches (No Gen.)' applied all the patterns leaned by our system for a given label and labeled every extraction as positive for the label. 'Pattern Matches' is similar to 'Pattern Matches (No Gen.)' except it used the dictionaries for generalizing the context, which increased the recall. `Pattern Matches in Dictionary' is the most conservative approach in which a token was labeled as positive only if it matched both by the dictionary and the learned patterns. That is, it filtered the output of 'Pattern Matching' to all the phrases that were also labeled by the dictionaries. All the pattern matching approaches have very low recall because many correct tokens did not occur in the patterns' context. 'Pattern Matches in Dictionary' has high precision because it is the most restricted approach of all, but suffers from low recall.

**Table 1: Precision, Recall, and $F_1$ scores of systems that use pattern matching when labeling data and our system.**

| | | | Asthma | | | |
|---|---|---|---|---|---|---|
| | DT | | | SC | | |
| System | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| Our system | **86.88** | **58.67** | **70.04** | 78.10 | **75.56** | **76.81** |
| Pattern Matches (No Gen.) | 45.26 | 13.73 | 21.07 | 50.75 | 12.33 | 19.85 |
| Pattern Matches | 36.58 | 16.60 | 22.84 | 43.07 | 17.10 | 24.48 |
| Patterns Matches in Dictionary | 80 | 13.28 | 22.78 | **83.51** | 15.47 | 26.11 |
| | | | **ENT** | | | |
| | **DT** | | | **SC** | | |
| System | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| Our system | 82.35 | **45.90** | **58.94** | 71.65 | **61.45** | **66.16** |
| Pattern Matches (No Gen.) | 47.36 | 4.71 | 8.57 | 46.15 | 0.98 | 1.92 |
| Pattern Matches | 40 | 6.55 | 11.26 | 53.12 | 8.85 | 15.17 |
| Pattern Matches in Dictionary | **90.90** | 5.46 | 10.30 | **94.11** | 8.33 | 15.31 |

### Experiment 2: Manually removing top negative words from MetaMap and OBA

We sorted all words extracted by MetaMap and OBA by their frequency and manually identified top 5 words that the authors judged as incorrect (we did not look at the context). We ran experiments in which those words were not labeled by OBA-C and MetaMap-C (marked as 'OBA-C-T5' and 'MetaMap-C-T5', respectively, in Table 2 and 3), that is, we added them to the stop words list. The motivation to compare the performance of these systems is when a user might be interested in manually identifying the top negative words and adding them to the stop words list. Removing the manually identified words generally increased precision, but reduced recall. We suspect the recall dropped because the words might be correct when they appeared in some contexts. The reason for the same scores for MetaMap-C and MetaMap-C-T5 for the SC label on the Asthma forum is that the negative words were already in the GoogleCommonList.

**Table 2: Precision, Recall, and $F_1$ scores of OBA when words in GoogleCommonList are not labeled ('-C' suffix), and when words in GoogleCommonList and in manually identified negative phrases are not labeled ('-C-T5' suffix).**

| | Asthma | | | | | |
|---|---|---|---|---|---|---|
| | DT | | | SC | | |
| System | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| OBA | 52.25 | **56.50** | 54.25 | 78.87 | **60.08** | 68.20 |
| OBA-C | 62.06 | 53.15 | 57.25 | 83.62 | 58.24 | **68.66** |
| OBA-C-T5 | **64.67** | 52.02 | **57.66** | **85.01** | 56.61 | 67.97 |
| | ENT | | | | | |
| | DT | | | SC | | |
| System | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| OBA | 43.22 | **55.73** | 48.68 | 67.51 | **50.52** | **57.59** |
| OBA-C | 49.73 | 51.36 | 50.53 | 70.55 | 46.18 | 55.82 |
| OBA-C-T5 | **53.71** | 51.36 | **52.51** | **82.31** | 42.01 | 55.63 |

**Table 3: Precision, Recall, and $F_1$ scores of MetaMap when words in GoogleCommonList are not labeled ('-C' suffix), and when words in GoogleCommonList and in manually identified negative phrases are not labeled ('-C-T5' suffix).**

| | Asthma | | | | | |
|---|---|---|---|---|---|---|
| | DT | | | SC | | |
| System | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| MetaMap | 68.42 | **57.56** | 62.52 | 58.63 | **80.24** | 67.75 |
| MetaMap-C | 77.60 | 54.98 | **64.36** | **70.28** | 75.15 | **72.63** |
| MetaMap-C-T5 | **78.68** | 53.13 | 63.43 | **70.28** | 75.15 | **72.63** |
| | ENT | | | | | |
| | DT | | | SC | | |
| System | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| MetaMap | 56.39 | **53.00** | 54.64 | 57.01 | **64.23** | 60.40 |
| MetaMap-C | 64.08 | 49.72 | **55.99** | 67.40 | 58.50 | 62.63 |
| MetaMap-C-T5 | **64.17** | 46.99 | 54.25 | **70.44** | 57.11 | **63.08** |

## ADDITIONAL METRIC

In our paper, the precision, recall, and F1 scores were calculated at the token level. Another way to calculate the scores is at the entity level, in which a phrase is considered positive only when all the words in the phrase are labeled positive. For example, if an entity is "salbutamol inhaler" and a system labels only “inhaler” as DT, then for the label DT, the entity-level number of true positives is 0, false negatives is 1, and false positives is 1. On the other hand, token-level number of true positives is 1, false negative is 1, and false positive is 0. Entity-level evaluation is preferred over token-level evaluation when extracting all words of an entity is more important than extracting parts of entity phrases. We present the results at the token level because for our task, identifying partial tokens in an entity (that is, “inhaler” in “salbutamol inhaler”) is still useful. Entity-level evaluation is commonly used for recognizing named entities, where, for example, the distinction between “Washington” and “Washington D. C.” is more prominent. Note that sometimes extracting partial phrases in our task will also lead to wrong number of token-level true positives (for example, extracting just “looking” in “trouble looking straight ahead”), but we did not observe it often in our experiments. For completeness, we present the performance of systems measured by entity-level precision, recall, and $F_1$ metrics in Table 4. Scores of all systems are better when measured at the token level than at the entity level because they get credit for extracting partial entities. The entity-level evaluation results show a similar trend as the token-level evaluation: our system performs better than other systems, albeit the difference is smaller for ENT-DT.

**Table 4: Entity-level Precision (Precision-E), Recall (Recall-E), and $F_1$ ($F_1$-E) scores of the systems.**

| | | | Asthma | | | |
|---|---|---|---|---|---|---|
| | | DT | | | SC | |
| System | Precision-E | Recall-E | $F_1$ -E | Precision-E | Recall-E | $F_1$ –E |
| OBA | 46.64 | 58.12 | 51.75 | 70.60 | 57.25 | 63.23 |
| OBA-C | 54.80 | 56.15 | 55.47 | **73.12** | 55.69 | 63.23 |
| MetaMap | 54.50 | 56.65 | 55.55 | 52.90 | **75.38** | 62.17 |
| MetaMap-C | 61.87 | 55.17 | 58.33 | 61.71 | 70.98 | 66.02 |
| Dictionary-F-C | 70.28 | 47.48 | 56.89 | 71.54 | 68.39 | 69.93 |
| Xu et al.-25 | **73.33** | 54.18 | 62.32 | 70.15 | 69.43 | 69.79 |
| Xu et al.-50 | 70.44 | 55.17 | 61.87 | 69.21 | 70.46 | 69.83 |
| CRF | 68.49 | 49.26 | 57.30 | 71.05 | 71.24 | 71.15 |
| CRF-2 | 69.65 | 49.75 | 58.04 | 70.43 | 70.98 | 70.70 |
| CRF-20 | 69.17 | 49.75 | 57.87 | 69.36 | 70.98 | 70.16 |
| Our system | 73.00 | **58.62** | **65.02** | 71.28 | 73.31 | **72.28** |
| | | | ENT | | | |
| | | DT | | | SC | |
| System | Precision-E | Recall-E | $F_1$ –E | Precision-E | Recall-E | $F_1$ -E |
| OBA | 29.79 | **49.57** | 37.22 | 56.57 | 44.79 | 50 |
| OBA-C | 34.16 | 46.21 | 39.28 | 56.78 | 40.72 | 47.43 |
| MetaMap | 40.97 | **49.57** | 44.86 | 48.51 | **59.04** | 53.26 |
| MetaMap-C | 46.28 | 47.05 | 46.66 | 56.10 | 54.07 | 55.06 |
| Dictionary-F-C | **66.23** | 42.85 | 52.04 | **65.57** | 50 | 56.73 |
| Xu et al.-25 | 60.71 | 42.85 | 50.23 | 64.63 | 50.45 | 56.67 |
| Xu et al.-50 | 47.66 | 42.85 | 45.12 | 64.78 | 52.03 | 57.71 |
| CRF | 62.65 | 43.69 | 51.48 | 63.27 | 50.67 | 56.28 |
| CRF-2 | 60.91 | 44.53 | 51.45 | 62.01 | 50.22 | 55.49 |
| CRF-20 | 51.45 | 44.53 | 47.74 | 61.38 | 50 | 55.11 |
| Our system | 63.52 | 45.37 | **52.94** | 62.53 | 55.88 | **59.02** |

## PARAMETER TUNING

In our paper, we tuned the parameters, such as *N*, *K*, and *T*, on the Asthma forum. In this section, we discuss the effect of varying some of the parameters (keeping others same as the final system) on extracting DT phrases from the Asthma forum. We experienced a similar effect of varying the parameters for extracting SC phrases from the Asthma forum.

### Phrase and pattern thresholds

Tables 5 and 6 show scores of our system when different phrase and pattern thresholds are used. In both cases, generally increasing the threshold resulted in higher precision but lower recall.

**Table 5: Scores when our system is run with different phrase threshold values. Increasing the threshold increases the precision but reduces recall. The value in bold was used in our final system.**

| Phrase threshold | Precision | Recall | F1 |
|---|---|---|---|
| 0.01 | 86.78 | 55.71 | 67.86 |
| 0.1 | 87.71 | 55.35 | 67.87 |
| **0.2** | 87.77 | **58.30** | **70.06** |
| 0.8 | **90.53** | 56.45 | 69.54 |
| 1.0 | **90.53** | 56.45 | 69.54 |

**Table 6: Scores when our system is run with different pattern threshold values. All other parameters remain unchanged. The threshold of 0.2 and 0.5 did not make a difference because all patterns extracted had score of more than 0.5. The threshold of 0.8 led to higher precision but lower recall. Threshold of 1.0 did not extract any patterns. The value in bold was used in our final system.**

| Pattern threshold | Precision | Recall | F1 |
|---|---|---|---|
| 0.2 | 87.77 | **58.30** | **70.06** |
| **0.5** | 87.77 | **58.30** | **70.06** |
| 0.8 | **90.68** | 53.87 | 67.59 |
| 1.0 | 89.65 | 47.97 | 62.50 |

### Number of phrases in each iteration (*N*)

In our system, we learned a maximum of 200 phrases (with maximum number of phrases in each iteration $N$=10 and maximum number of iterations $T$=20). Table 7 shows scores for different combinations of values of $N$ and $T$, keeping the total number of phrases learned constant.

**Table 7: Scores when our system is run with different values of *N* and *T*. The values in bold were used in our final system.**

| N | T | Precision | Recall | F1 |
|---|---|---|---|---|
| 5 | 40 | 86.41 | **58.67** | 69.89 |
| **10** | **20** | 87.77 | 58.30 | **70.06** |
| 40 | 5 | **89.22** | 54.98 | 68.03 |

### Number of patterns in each iteration (*K*)

Table 8 shows results for different values of K, that is, the maximum number of pattern learned in each iteration.

**Table 8: Scores when our system is run with different values of *K*. Increasing *K* decreases precision but improves recall. The values shown in bold were used in our final system.**

| K | T | Precision | Recall | F1 |
|---|---|---|---|---|
| 20 | 20 | **88.75** | 55.35 | 68.18 |
| **50** | **20** | 87.77 | 58.30 | **70.06** |
| 100 | 20 | 86.41 | **58.67** | 69.89 |

## ANECDOTAL EFFICACY LABELS

In the paper, we demonstrated a use case of the system to explore alternative treatments people use for a symptom or condition. We manually labeled posts that mentioned a particular treatment with the following labels.

Strongly positive: The person has explicitly mentioned that the treatment is helping the subject of the post (many times the posts discuss health of a family member) for Diabetes. Example: "… *A relative with the same problem told her about taking cinnamon gel tabs which had greatly helped her. She found a brand at the local health store by the name of NewChapter titled Cinnamon Force. She was afraid to take it with so many other medications and it sat in the cabinet about five months. Last week, she got brave and took two tabs behind the two largest meals of the day.Wow! the level dropped down into the safe range and has remained there for several days.All that I can tell you about the product, is that it contains 140mg of cinnamon per gel tab.We are so thrilled that after so many years of frustration, that we see a great change in blood sugar levels…"*

Weakly positive: The subject of the post is either using the treatment or heard/read positive effects of the treatment for Diabetes. Example: "… *Some people do think things such as vinegar help. My belief is those things are worth trying but they are secondary to tried and true things such as weight loss, exercise and lowering carb intake.*"

Neutral: The subject of the post is neither using the treatment nor expressed any sentiment about it in the post. Example: "… *I may be wrong, but I haven't heard of cinnamon lowering glucose levels. Please take your mother to a doctor for a checkup asap*…" Posts that asked a question about using the about treatment were also labeled neutral. For example, "*Does vinegar help diabetes*".

Weakly negative: The post mentioned that the user has heard that the treatment does not work. For example, people citing studies that showed inconclusive evidence of the efficacy of the treatment. Example: "… *Studies now show that cinnamon doesn't lower glucose levels, but has been known to regulate blood pressure. I can vouch for the latter*…"

Strongly negative: The post mentioned that the treatment is not working from personal experience of the subject of the post (for example, a family member). Example: "*I have tried the Apple Cider Vinegar and it didn't work for me* …"

## STATISTICAL SIGNIFICANCE TESTING

We tested the statistical significance of the improvement of our system over a baseline using approximate randomization[1,2] implemented by SIGFv2[3] , commonly used for statistical significance testing for named entity recognition systems.  It does not assume that the model is representative. We assumed each token to be an observation and randomized the observations 10,000 times.

**References**

[1] E. Noreen. Computer-intensive Methods for Testing Hypotheses: An Introduction. John Wiley and Sons Inc. 1989.

[2] A. Yeh. More accurate tests for the statistical significance of result differences. In the Proceedings of the International Conference on Computational Linguistics. 2000. p. 947--953.

[3] Sebastian Padó. User's guide to SIGF: Significance testing by approximate randomization. 2006