Twitter社　山本様、藤井様

楽天株式会社楽天技術研究所の村上と申します。この度は非平常時の言語情報としての御社のTweetの利用について、ご協力のご相談をさせて頂きたく思います。

私は通常、企業にて（言語を機械に処理させる）自然言語処理分野の研究開発を行っています。昨年3月の東日本震災時に、企業や大学の垣根を越えて自然言語処理に関わる研究者や技術者およそ70名が協力して「ANPI\_NLP」という名前の有志ワーキンググループを立ち上げ、我々が持つ言語処理技術を利用してインターネット上に溢れる被災者の情報を集約、整理するプロジェクトを進めました。

特に注力したのが安否情報抽出です。Mixi内の掲示板、ブログ等のCGMにも着目しましたが、最も注力したのは即時性の高かった御社のTweetでした。私達は前もってクロールした数万のTweetから被災者の安否情報や避難所の情報を分類し、被災者名や住所情報などを抽出したのち、Google社が提供していたGoogle Person FinderのDBと照合、更新するという活動を行いました。

　2012年3月13日から開かれた第18回言語処理学会年次大会にて、「災害とNLP」というテーマセッションを行いました。ここでは昨年行った活動、引き続き行われている研究発表などのほかに、我々言語処理に携わる者が、再び起こる災害時を想定したタスクに今後、どのように取り組んでいくかについて議論しました。

　私達は災害NLPという一つの実質的活動を続けていくつもりなのですが、言語処理の研究開発には言語データが必須であり、昨年の活動でも様々な種類のデータを利用しました。その中でも御社のTweetには最も注力し、61,375のTweetを収集、33,000程のTweetに対して「どのような情報を含んでいるか」を示すトピックタグと、Tweet中の「人名」や「地名」に対してのタグを付与することでデータを作成しました。このTweetデータは他のものに比べても様々な言語表現、言語現象が含まれており、極めて貴重なデータであると考えています。私達は研究開発を定常的に進めていくために、このタグを付けたTweetを、改めて研究目的において研究機関で、共有、利用したいと考えています。

そこで、この私達がタグを付与したデータについてご相談させてください。タグ付けを行った実際のTweetの例を示します。下線部の＜＞に囲まれた部分が付与したタグです。

（例）

<location>仙台</location>に住む<person>日本圭子</person>、<person>日本ゆうき</person>を探しています。どんな情報でもかまいませんので情報提供お願いします。<L>

私達の勝手な解釈で話を進めさせて頂きます。データを公開および研究者間での共有には２つの乗り越えなければいけない問題があります。１つは「個人情報保護法」、そしてもう１つは御社Twitter社の所有するTweetの利用に関する規約です。

　個人情報保護法について簡単に調べたところ、情報の利用は非常時においてはその適用範囲ではないとのことで、私達が個人情報の固まりであるTweet群を活動のために利用した事そのものは問題がないと理解しています。しかし震災から1年以上経った今、平常時に戻っており、法の範疇にあります。いかなる形態にせよ、データの公開共有には個人情報を予め排除する必要があります。

　ただ、個人情報保護法への対応については、言語処理分野はある程度の知識と経験を有しており、過去にも個人情報の多く含まれるデータを扱っています。私達の経験では、個人の特定が可能となる情報である「人名」、「地名（住所を含む）」「電話番号」などが個人情報として該当することから、これらをアルファベットの並び等でマスクするなど、個人の特定が不可能になる形状に予め処理を施すことで法を回避できると考えています。私達言語処理の分野では「人名」「地名」を適切に認識する「固有表現抽出」というタスクがあり、可能であればマスクを施さない生データの利用がもっとも理想的ではありますが、それらの固有表現を認識するための周辺の文脈情報が利用できるのであれば、それほど大きな影響はありません。

　一方、御社Twitter社の第３者によるAPI等を用いたTweetの（再）利用に関する規約も重要な検討事項になります。不勉強ではありますが調べたところによると、第３者がTweetを取得する場合Twitter社の規約を遵守するという前提であること、Tweetの著者の要請等によりそのTweetの削除の必要がある時は速やかに全ての取得者はそれに従う必要があること、は理解できました。規約を遵守するには、例えば私個人、もしくは有志のグループがタグ付きTweetデータを研究者に配布する際は、削除等の義務を負うこととなりますが配布先での削除の確認は現実的には難しいと考えています。

　そこで私達として現実的に可能なデータ共有の手段は、私個人、もしくは有志のグループが配布元となるのではなく、データを必要とする研究者側でTwitter社の規約、及び個人情報保護法の責務を負う形ではどうかと考えています。具体的には以下のような方法です。

1. Tweetそのものではなく、必要なTweetのIDを私個人もしくはグループが公開
2. 研究者側が個人でTweetを取得
3. IDと同時に、Tweet中の人名、地名、トピックタグの位置情報を配布

　ただこの方法ですと、恐らくデータを利用する研究機関数は多くて２０−３０ではありますがTwitter社へのトラフィックが増える点、個人を特定できる情報を研究者側で所有してしまうという点で個人情報保護法に抵触するかもしれない、の２点の懸念事項が残ります。

この方法に対して、個人情報保護法への対応を主眼に置いた方法としては以下の方法が考えられます。

1. 有志のグループ側で全てのTweetに対してタグ付け処理
2. 個人情報保護法に抵触する情報に対してはアルファベット並び等のマスク処理

これに加えて、もし可能であるなら

（３）タグ付きデータの管理をTwitter社に委託し、データ配布元になって頂く

ここに関しては種々の難しさがあると思っていますが、もし委託可能であるのであれば、データそのものの権利の委譲まであり得ると考えています。

　私達はこの次に再び起こるであろう何らかの災害に対して、個人が発信する情報を捉えて必要とされている情報を何らかの形で社会と共有するという目的のための言語処理技術の研究開発を継続的に行いたいと考えています。そのためには言語データの共有は必須であり、データ元であるTwitter社様に我々の目的と活動を知って頂き、この度作成したタグ付きTweetデータを研究機関で利用可能にするためにお力を貸して頂けないかご相談させて頂きました。ご質問、ご意見等ありましたら、是非お願い致します。もしデータ公開共有に向けて私たちの方ですべきことがありましたらできるだけの事をする所存です。

どうぞよろしくお願い致します。