Sentence Simplification and Its Application to Information Extraction

Yifan Peng\textsuperscript{1}, Cathy Wu\textsuperscript{1,2}, and Vijay Shanker\textsuperscript{1}

\textsuperscript{1}Computer and Information Sciences, University of Delaware
\textsuperscript{2}Center for Bioinformatics & Computing Biology, University of Delaware

\section*{Motivation}
Long and complicated sentences pose challenges to information extraction systems. In this work, we explore methods to transform complex sentences into sentences that are simpler in grammatical structure. The simplified sentences are intended to be input into information extraction systems. We expect it will improve the accuracy of the system, but not yet evaluate it.

\section*{Relation Extraction}
Relation extraction is the task of finding a relevant semantic relation between one or more given target entities (usually two) from text (usually in single sentence).

\begin{itemize}
  \item PKC phosphorylates GAP-43 on serine 41.
\end{itemize}

Pattern: \textit{A phosphorylates B}, when \textit{A} and \textit{B} are proteins.

\section*{Relation Extraction with Simplified Sentences}

\begin{itemize}
  \item \textbf{Appositive, Conjunction, and Relative Clause}
    \begin{itemize}
      \item \textbf{Appositive}
        \begin{itemize}
            \item Active Raf-1 phosphorylates and activates MEK1, which in turn phosphorylates and activates the MAP kinases/extracellular signal regulated kinases, \textit{ERK1 and ERK2}.
        \end{itemize}
      \item \textbf{Conjunction}
        \begin{itemize}
            \item Active Raf-1 phosphorylates and activates MEK1, which in turn phosphorylates and activates the MAP kinases/extracellular signal regulated kinases, \textit{ERK1 and ERK2}.
        \end{itemize}
      \item \textbf{Relative clause}
        \begin{itemize}
            \item Active Raf-1 phosphorylates and activates MEK1, which in turn phosphorylates and activates the MAP kinases/extracellular signal regulated kinases, \textit{ERK1 and ERK2}.
        \end{itemize}
    \end{itemize}
  \item \textbf{Complex sentence with complicated rules}
    \begin{itemize}
      \item 1. Active Raf-1 phosphorylates MEK1.
      \item 2. MEK1 phosphorylates ERK1.
      \item 3. MEK1 phosphorylates ERK2.
    \end{itemize}
  \item \textbf{Simplified sentences with simple rules}
    \begin{itemize}
      \item Active Raf-1 phosphorylates and activates MEK1, which in turn phosphorylates and activates the MAP kinases/extracellular signal regulated kinases, \textit{ERK1 and ERK2}.
    \end{itemize}
\end{itemize}

\section*{Type of Sentence Simplification}

\begin{itemize}
  \item \textbf{Appositive}
    \begin{itemize}
        \item Active Raf-1 phosphorylates and activates MEK1, which in turn phosphorylates and activates the MAP kinases/extracellular signal regulated kinases, \textit{ERK1 and ERK2}.
    \end{itemize}
  \item \textbf{Conjunction}
    \begin{itemize}
        \item Active Raf-1 phosphorylates and activates MEK1, which in turn phosphorylates and activates the MAP kinases/extracellular signal regulated kinases, \textit{ERK1 and ERK2}.
    \end{itemize}
  \item \textbf{Relative clause}
    \begin{itemize}
        \item Active Raf-1 phosphorylates and activates MEK1, which in turn phosphorylates and activates the MAP kinases/extracellular signal regulated kinases, \textit{ERK1 and ERK2}.
    \end{itemize}
\end{itemize}

\section*{Work Flow}

\begin{itemize}
  \item \textbf{Sentences}
        \begin{itemize}
            \item Part-of-speech tagger
            \item Named-entity recognizer
            \item Chunker
        \end{itemize}
  \item \textbf{Tokenization}
    \begin{itemize}
        \item Pre-processing
            \begin{itemize}
                \item Coordination rules
                \item Relative clause rules
                \item Appositive rules
            \end{itemize}
    \end{itemize}
  \item \textbf{Sentence Simplification}
    \begin{itemize}
        \item Sentences
        \item Syntactic description
            \begin{itemize}
                \item Mark-up
            \end{itemize}
    \end{itemize}
\end{itemize}

\section*{Intergrate INTO Information Extraction}

\begin{itemize}
  \item \textbf{Feed existing Relation Extraction tools with simplified sentences.}
    \begin{itemize}
      \item Immunoblotting with phospho-specific antibodies confirmed cAMP/vasopressin-dependent phosphorylation at Thr96, Ser126, and Ser874 of the cotransporter NKCC2 and at Ser552 of beta-catenin.
      \item RLIMS-P\textsuperscript{50} did not extract phosphorylation at Ser552 of beta-catenin;
      \item RLIMS-P extracted both phosphorylation correctly on simplified sentences;
      \item We will be evaluating RLIMS-P performance on simplified vs original sentences.
    \end{itemize}
  \item \textbf{Use simplification during learning.}
  \begin{itemize}
    \item Machine learning algorithm
      \begin{itemize}
        \item Learning
          \begin{itemize}
            \item Original annotated corpus
            \item Simplified annotated corpus
            \item Learnt model
          \end{itemize}
        \item Application
          \begin{itemize}
            \item Original test text
            \item Simplified test text
            \item Result
          \end{itemize}
      \end{itemize}
  \end{itemize}
\end{itemize}