大規模コーパスを扱うためのツール群

岡野原　大輔　辻井　潤一　東京大学

- Google Nグラムコーパスや Wikipediaなど大規模コーパスを利用した自然言語処理を支援するためのツールを開発した
- tx：木の簡潔表現を利用したtrie
- bep：最小完全ハッシュ関数による連想配列
- oll：複数のオンライン学習手法を備えた学習器
- データ構造や学習手法の最新の研究成果を取り入れ自然言語処理向けに最適化
- 全て修正BSDライセンスでダウンロード可能
  - 複数の企業や研究所での利用実績有

- tx
  - 大規模キー集合(数百億キー）に対し共通接頭辞検索や連想配列をサポート
  - キー自身の約半分のメモリー1300万キーで約50MB
  - 筑波エクスプレス車内で開発された
- bep
  - 連想配列のみだが1キーあたり約4bit
  - 別府温泉の開発合宿で開発された
- oll
  - 6種類のオンライン学習手法をサポート
  - 1訓練例あたり約3μ秒
  - 論文締切間際の逃避行動で開発された

こんなツール欲しいという要望も受け付けます！