Przetwarzanie języka polskiego za pomocą kategorialnej gramatyki logicznej

Wojciech Jaworski

Instytut Informatyki
Uniwersytet Warszawski

12 kwietnia 2013

Wprowadzenie

- Systemy gramatyczne oparte na logice liniowej (Type Logical Grammar, czy Combinatory Categorial Grammar) są z powodzeniem stosowane do opisu i przetwarzania języków naturalnych.
- Źródłem ich użyteczności jest fakt, że języki naturalne mają własność logicznej liniowości:
  - każde słowo w zdaniu jest wykorzystane dokładnie raz w drzewie rozbioru składniowego, a jego znaczenie występuje dokładnie raz w formie logicznej tego zdania;
  - spójniki występujące w logice liniowej w naturalny sposób wyrażają pojęcia używane przy opisie języka;
  - drzewa rozbioru gramatycznego są dowodami w tej logice.

Typy

- W wyniku segmentacji zdanie zostaje zamienione na ciąg tokenów $\gamma_1, \gamma_2, \ldots, \gamma_n$,
- a w wyniku analizy morfologicznej każdemu tokenowi zostaje przypisana formuła (typ) $\varphi_1, \ldots, \varphi_n$.
- na przykład zdanie „Jan widzi stół.” zostanie podzielone na tokeny:

$$\text{Jan}_1,\ \text{widzi}_2,\ \text{stół}_3, .{}_4$$

którym przyporządkowane zostaną typy

$$\mathrm{NP},\ (\mathrm{VP} \setminus \mathrm{NP}) / \mathrm{NP},\ \mathrm{NP},\ \mathrm{S} \setminus \mathrm{VP}$$

- Typy składają się z symboli atomowych reprezentujących rodzaj frazy (np. NP, VP) i kategorie gramatyczne takie jak przypadek, liczba, rodzaj czy osoba (np. $\mathrm{nom}, \mathrm{sg}, \mathrm{m}_3$, sec)
- Symbole te łączone są za pomocą spójników logicznych.

Wywód gramatyczny

- Drzewem wyprowadzenia (wywodu gramatycznego) jest dla nas dowód w intuicjonistycznej niekomutatywnej logice liniowej.
- Założeniami w dowodzie są dla nas sekwenty utworzone z wyodrębnionych w wyniku segmentacji tokenów i przypisanych im typów:

$$\gamma_1 \vdash \varphi_1, \quad \gamma_2 \vdash \varphi_2, \quad \ldots \quad , \gamma_n \vdash \varphi_n.$$

- Tezą w dowodzie jest sekwent

$$\gamma_1, \gamma_2, \ldots, \gamma_n \vdash \mathrm{S}$$

oznaczający, że z ciągu złożonego ze wszystkich tokenów potrafimy wywieść całe zdanie.

Przykładowy wywód gramatyczny

- Będziemy korzystać z aksjomatów

$$\text{Jan}_1 \vdash \text{NP},\ \text{widzi}_2 \vdash (\text{VP} \setminus \text{NP}) / \text{NP},\ \text{stół}_3 \vdash \text{NP},\ .\!_4 \vdash \text{S} \setminus \text{VP}$$

- oraz z reguł wnioskowania *modus ponens*:

$$\frac{\Gamma \vdash \psi / \varphi \qquad \Delta \vdash \varphi}{\Gamma, \Delta \vdash \psi}\,[/E] \qquad \frac{\Delta \vdash \varphi \qquad \Gamma \vdash \psi \setminus \varphi}{\Delta, \Gamma \vdash \psi}\,[\setminus E]$$

- Otrzymamy następujące drzewo wywodu:

$$\dfrac{\dfrac{\text{Jan}_1 \vdash \text{NP} \qquad \dfrac{\text{widzi}_2 \vdash (\text{VP} \setminus \text{NP}) / \text{NP} \qquad \text{stół}_3 \vdash \text{NP}}{\text{widzi}_2, \text{stół}_3 \vdash \text{VP} \setminus \text{NP}}}{\text{Jan}_1, \text{widzi}_2, \text{stół}_3 \vdash \text{VP}} \qquad .\!_4 \vdash \text{S} \setminus \text{VP}}{\text{Jan}_1, \text{widzi}_2, \text{stół}_3, .\!_4 \vdash \text{S}}$$

Spójniki: tensor, wraz, top

- Tensor $\varphi \bullet \psi$ pozwala zapisać wektory kategorii gramatycznych oraz konkatenację fraz.

$$\text{Jan} \vdash \text{NP} \bullet \text{sg} \bullet \text{nom} \bullet \text{m}_1 \bullet \text{ter}$$

$$\frac{\Gamma \vdash \varphi \bullet \psi \qquad \Delta, \varphi, \psi, \Delta' \vdash \sigma}{\Delta, \Gamma, \Delta' \vdash \sigma}\,[\bullet E] \qquad \frac{\Gamma \vdash \psi \qquad \Delta \vdash \varphi}{\Gamma, \Delta \vdash \psi \bullet \varphi}\,[\bullet I]$$

- Wraz $\varphi \,\&\, \psi$ reprezentuje niejednoznaczność, prawo wyboru między typami.

$$\text{Jana} \vdash \text{NP} \bullet \text{sg} \bullet (\text{gen} \,\&\, \text{acc}) \bullet \text{m}_1 \bullet \text{ter}$$

$$\frac{\Gamma \vdash \psi \,\&\, \varphi}{\Gamma \vdash \psi}\,[\&E] \quad \frac{\Gamma \vdash \psi \,\&\, \varphi}{\Gamma \vdash \varphi} \qquad \frac{\Gamma \vdash \psi \qquad \Gamma \vdash \varphi}{\Gamma \vdash \psi \,\&\, \varphi}\,[\&I]$$

- Stała top $\top$ to uogólnienie wszystkich typów atomowych.

$$\text{widzi}_i \vdash (\text{VP} \setminus \text{NP} \bullet \text{sg} \bullet \text{nom} \bullet \top) \,/\, \text{NP} \bullet \top \bullet \text{acc} \bullet \top$$

$$\gamma \vdash \top\,[\top I]$$

- Jako argument implikacji top pozwala reprezentować polimorfizm.

Przykład

$$\frac{\Gamma \vdash \varphi \bullet \psi \qquad \Delta, \varphi, \psi, \Delta' \vdash \sigma}{\Delta, \Gamma, \Delta' \vdash \sigma}\,[\bullet E] \qquad \frac{\Gamma \vdash \psi \qquad \Delta \vdash \varphi}{\Gamma, \Delta \vdash \psi \bullet \varphi}\,[\bullet I]$$

$$\frac{\Gamma \vdash \psi \,\&\, \varphi}{\Gamma \vdash \psi}\,[\&E] \quad \frac{\Gamma \vdash \psi \,\&\, \varphi}{\Gamma \vdash \varphi} \qquad \frac{\Gamma \vdash \psi \qquad \Gamma \vdash \varphi}{\Gamma \vdash \psi \,\&\, \varphi}\,[\&I]$$

$$\gamma \vdash \top\,[\top I] \qquad \varphi \vdash \varphi\,[\mathit{Axiom}]$$

$$\frac{\mathrm{stół}_3 \vdash \mathrm{NP} \bullet \mathrm{sg} \bullet (\mathrm{nom} \,\&\, \mathrm{acc}) \qquad \dfrac{\mathrm{NP} \vdash \mathrm{NP} \qquad \mathrm{sg} \vdash \top \qquad \dfrac{\mathrm{nom} \,\&\, \mathrm{acc} \vdash \mathrm{nom} \,\&\, \mathrm{acc}}{\mathrm{nom} \,\&\, \mathrm{acc} \vdash \mathrm{acc}}}{\mathrm{NP}, \mathrm{sg}, \mathrm{nom} \,\&\, \mathrm{acc} \vdash \mathrm{NP} \bullet \top \bullet \mathrm{acc}}}{\mathrm{stół}_3 \vdash \mathrm{NP} \bullet \top \bullet \mathrm{acc}}$$

$$\frac{\mathrm{widzi}_2 \vdash (\mathrm{VP} \setminus \mathrm{NP} \bullet \mathrm{sg} \bullet \mathrm{nom}) / \mathrm{NP} \bullet \top \bullet \mathrm{acc} \qquad \dfrac{\mathrm{stół}_3 \vdash \mathrm{NP} \bullet \mathrm{sg} \bullet (\mathrm{nom} \,\&\, \mathrm{acc})}{\mathrm{stół}_3 \vdash \mathrm{NP} \bullet \top \bullet \mathrm{acc}}\,*}{\mathrm{widzi}_2, \mathrm{stół}_3 \vdash \mathrm{VP} \setminus \mathrm{NP} \bullet \mathrm{sg} \bullet \mathrm{nom}}$$

Spójnik plus

- Plus $\varphi \oplus \psi$ jest to uogólnienie typu φ oraz ψ, czyli jeden z tych typów ze wskaźnikiem, o który typ faktycznie chodzi.

$$\text{jest}_i \vdash \text{VP} \,/\, ((\text{NP} \bullet \text{inst}) \oplus (\text{AdjP} \bullet \text{nom}))$$

- Jako argument implikacji, plus pozwala reprezentować polimorficzne funktory.

$$\frac{\Gamma \vdash \psi \oplus \varphi \qquad \Delta, \psi, \Delta' \vdash \sigma \qquad \Delta, \varphi, \Delta' \vdash \sigma}{\Delta, \Gamma, \Delta' \vdash \sigma}\,[\oplus E]$$

$$\frac{\Gamma \vdash \psi}{\Gamma \vdash \psi \oplus \varphi}\,[\oplus I] \qquad \frac{\Gamma \vdash \varphi}{\Gamma \vdash \psi \oplus \varphi}$$

Stała jeden

- Stała jeden 1 reprezentuje białe znaki.
- Jako argument implikacji jedynka użyta z plusem tworzy argumenty opcjonalne.

$$\mathrm{widzi}_i \vdash (\mathrm{VP} \setminus ((\mathrm{NP} \bullet \mathrm{sg} \bullet \mathrm{nom} \bullet \top) \oplus 1)) / ((\mathrm{NP} \bullet \top \bullet \mathrm{acc} \bullet \top) \oplus 1)$$

$$\vdash 1\ [1I] \qquad \frac{\Gamma \vdash \sigma \qquad \Delta \vdash 1}{\Gamma, \Delta \vdash \sigma}\ [1E]$$

- Przykład:

$$\frac{\Gamma \vdash \psi}{\Gamma \vdash \psi \oplus \varphi}\ [\oplus I] \quad \frac{\Gamma \vdash \varphi}{\Gamma \vdash \psi \oplus \varphi}$$

$$\frac{\mathrm{widzi}_2 \vdash (\mathrm{VP} \setminus \mathrm{NP} \bullet \mathrm{sg}) / (\mathrm{NP} \bullet \top \oplus 1) \qquad \dfrac{\vdash 1}{\vdash \mathrm{NP} \bullet \top \oplus 1}}{\mathrm{widzi}_2 \vdash \mathrm{VP} \setminus \mathrm{NP} \bullet \mathrm{sg}}$$

Reprezentacja swobodnego szyku

- Zdanie

  Jan widzi stół

  możemy zapisać w dowolnej kolejności nie zmieniając zasadniczo jego znaczenia, np.:

  Jan stół widzi *widzi Jan stół* *stół widzi Jan*

- Aby wygodnie reprezentować swobodny szyk wprowadzam obustronną implikację:

$$\frac{\Gamma \vdash \psi \mid \varphi \qquad \Delta \vdash \varphi}{\Gamma, \Delta \vdash \psi}\,[\mid E]\,\frac{\Delta \vdash \varphi \qquad \Gamma \vdash \psi \mid \varphi}{\Delta, \Gamma \vdash \psi}$$

- oraz zbiory argumentów

$$\frac{\Gamma \vdash \psi\{\mid_1 \varphi_1, \mid_2 \varphi_2, \ldots, / \varphi_i, \ldots, \mid_n \varphi_n\} \qquad \Delta \vdash \varphi_i}{\Gamma, \Delta \vdash \psi\{\mid_1 \varphi_1, \mid_2 \varphi_2, \ldots, \mid_n \varphi_n\}}\,[\{\}E]$$

  oraz analogiczne reguły dla $\backslash$ i $\mid$.

$$\text{widział} \Longrightarrow \mathrm{VP}\{ \mid \mathrm{NP} \bullet \mathrm{sg} \bullet \mathrm{nom} \bullet (\mathrm{m}_1 \oplus \mathrm{m}_2 \oplus \mathrm{m}_3) \bullet \mathrm{ter}, \mid \mathrm{NP} \bullet \top \bullet \mathrm{acc} \bullet \top \bullet \top \}$$

Modyfikatory

- W zdaniu

  Jan widzieć musi stół

  czasownik *musi* łączy się z bezokolicznikiem *widzieć* zanim bezokolicznik pobierze swoje dopełnienie.

$$\dfrac{\dfrac{\mathrm{Jan}_1 \vdash \mathrm{nom} \qquad \dfrac{\mathrm{widzieć}_2 \vdash \mathrm{inf}\{|\,\mathrm{pro}, |\,\mathrm{acc}\} \qquad \mathrm{musi}_3 \vdash \mathrm{VP}\{|\,\mathrm{nom}\} \parallel \mathrm{inf}\{|\,\mathrm{pro}\}}{\mathrm{widzieć}_2, \mathrm{musi}_3 \vdash \mathrm{VP}\{|\,\mathrm{nom}, |\,\mathrm{acc}\}}}{\mathrm{Jan}_1, \mathrm{widzieć}_2, \mathrm{musi}_3 \vdash \mathrm{VP}\{|\,\mathrm{acc}\}} \qquad \mathrm{stół}_4 \vdash \mathrm{acc}}{\mathrm{Jan}_1, \mathrm{widzieć}_2, \mathrm{musi}_3, \mathrm{stół}_4 \vdash \mathrm{VP}}$$

- Wprowadzamy operatory $/\!\!/$, $\backslash\!\!\backslash$ i $\parallel$ reprezentujące modyfikatory

$$\frac{\Gamma \vdash \psi\{\Sigma\} /\!\!/ \varphi\{\Theta\} \qquad \Delta \vdash \varphi\{\Theta, \Pi\}}{\Gamma, \Delta \vdash \psi\{\Sigma, \Pi\}}\ [/\!\!/\, E]$$

$$\frac{\Delta \vdash \varphi\{\Theta, \Pi\} \qquad \Gamma \vdash \psi\{\Sigma\} \backslash\!\!\backslash \varphi\{\Theta\}}{\Delta, \Gamma \vdash \psi\{\Sigma, \Pi\}}\ [\backslash\!\!\backslash\, E]$$

Model dla fragmentu multiplikatywnego

- Niech $\mathcal{A}$ będzie zbiorem typów atomowych (np.: $\mathrm{NP}, \mathrm{VP}, \ldots$).
- Język formuł kategorialnych $\mathcal{F}$ definiujemy jako

$$\begin{aligned}\mathcal{F} \quad ::= \quad & \mathcal{A} \mid \mathcal{F} \bullet \mathcal{F} \mid \mathcal{F} \,/\, \mathcal{F} \mid \mathcal{F} \setminus \mathcal{F} \mid \mathcal{F} | \mathcal{F} \\ & \mid \mathcal{F}\{|_1 \, \mathcal{F}, \ldots, |_n \, \mathcal{F}\} \mid \mathcal{F} /\!\!/ \mathcal{F} \mid \mathcal{F} \setminus\!\!\setminus \mathcal{F} \mid \mathcal{F} \parallel \mathcal{F}\end{aligned}$$

- Rozważamy struktury $\mathcal{W} = \langle W, \cdot \rangle$.
- W rozumiemy jako zbiór *zasobów lingwistycznych*: tokenów (słów, znaków interpunkcyjnych, itp.) oraz ciągów tokenów.
- Operator '$\cdot$' jest operatorem konkatenacji.
- Zakładamy, że $\cdot$ jest łączny:

$$\forall x, y, z[(x \cdot y) \cdot z = x \cdot (y \cdot z)].$$

Wartościowanie

- Model uzyskujemy dodając wartościowanie v, które każdemu typowi atomowemu γ przypisuje podzbiór W

$$v(\gamma) \subseteq W \text{ dla } \gamma \in \mathcal{A}$$

- wartościowanie formuł złożonych definiujemy indukcyjnie:

$$\begin{aligned}
v(\varphi \bullet \psi) &= \{x \cdot y : x \in v(\varphi) \wedge y \in v(\psi)\} \\
v(\psi \,/\, \varphi) &\subseteq \{x : \forall y[y \in v(\varphi) \Rightarrow x \cdot y \in v(\psi)]\} \\
v(\psi \setminus \varphi) &\subseteq \{x : \forall y[y \in v(\varphi) \Rightarrow y \cdot x \in v(\psi)]\} \\
v(\varphi \mid \psi) &\subseteq \{x : \forall y[y \in v(\psi) \Rightarrow x \cdot y \in v(\varphi) \wedge y \cdot x \in v(\varphi)]\} \\
v(\varphi\{|_1\ \psi_1, \ldots, /\ \psi_i, \ldots, |_n\ \psi_n\}) &\subseteq \{x : \forall y[y \in v(\psi_i) \Rightarrow x \cdot y \in v(\varphi\{|_1\ \psi_1, \ldots, |_{i-1}\ \psi_{i-1}, |_{i+1}\ \psi_{i+1}, \ldots, |_n\ \psi_n\})]\} \\
v(\varphi\{|_1\ \psi_1, \ldots, \setminus\ \psi_i, \ldots, |_n\ \psi_n\}) &\subseteq \{x : \forall y[y \in v(\psi_i) \Rightarrow y \cdot x \in v(\varphi\{|_1\ \psi_1, \ldots, |_{i-1}\ \psi_{i-1}, |_{i+1}\ \psi_{i+1}, \ldots, |_n\ \psi_n\})]\} \\
v(\varphi\{|_1\ \psi_1, \ldots, |\ \psi_i, \ldots, |_n\ \psi_n\}) &\subseteq \{x : \forall y[y \in v(\psi_i) \Rightarrow x \cdot y \in v(\varphi\{|_1\ \psi_1, \ldots, |_{i-1}\ \psi_{i-1}, |_{i+1}\ \psi_{i+1}, \ldots, |_n\ \psi_n\}) \wedge \\
&\quad y \cdot x \in v(\varphi\{|_1\ \psi_1, \ldots, |_{i-1}\ \psi_{i-1}, |_{i+1}\ \psi_{i+1}, \ldots, |_n\ \psi_n\})]\}
\end{aligned}$$

Poprawność i pełność

Theorem

$\varphi_1, \ldots, \varphi_n \vdash \psi$ jest dowodliwe wtw. dla każdego wartościowania v na każdej strukturze $\mathcal{W}$ zachodzi $v(\varphi_1 \bullet \ldots \bullet \varphi_n) \subseteq v(\psi)$.

Dowód.

- Aby udowodnić pełność wprowadzamy następujący model kanoniczny $\langle W_K, \cdot_K, v_K \rangle$, gdzie
  1. W_K jest zbiorem formuł $\mathcal{F}$;
  2. $\sigma = \varphi \cdot_K \psi$ wtw. sekwent $\sigma \vdash \varphi \bullet \psi$ jest dowodliwy;
  3. $\varphi \in v_K(\psi)$ wtw. sekwent $\varphi \vdash \psi$ jest dowodliwy.
- W tym modelu, jeśli $\varphi \vdash \psi$ jest niedowodliwe to $\varphi \notin v_K(\psi)$.
- Jako, że $\varphi \in v_K(\varphi)$ otrzymujemy $v_K(\varphi) \not\subseteq v_K(\psi)$.

Lista nieobecności

- koordynacja, np. *Dajcie wina i całą świnię*;

$$\dfrac{\text{Dajcie} \vdash \mathrm{VP}\{\mid \mathrm{gen} \oplus \mathrm{acc}\} \qquad \dfrac{\dfrac{\text{wina} \vdash \mathrm{gen}}{\text{wina} \vdash \mathrm{gen} \oplus \mathrm{acc}} \quad \text{i} \vdash \mathrm{conj} \quad \dfrac{\text{całą świnię} \vdash \mathrm{acc}}{\text{całą świnię} \vdash \mathrm{gen} \oplus \mathrm{acc}}}{\text{wina i całą świnię} \vdash (\mathrm{gen} \oplus \mathrm{acc})^\star}}{\dfrac{\text{Dajcie wina i całą świnię} \vdash \mathrm{VP}\{\mid \mathrm{gen} \oplus \mathrm{acc}\} \bullet (\mathrm{gen} \oplus \mathrm{acc})^\star}{\dfrac{\text{Dajcie wina i całą świnię} \vdash \mathrm{VP}^\star}{\text{Dajcie wina i całą świnię} \vdash \mathrm{VP}}}}$$

- semantyka — liniowy rachunek lambda;

$$\text{widział} \vdash \mathrm{VP}\{\mid \mathrm{nom}, \mid \mathrm{acc}\} : \lambda s, o.\mathrm{widzi}(s, o) \qquad \dfrac{\Gamma \vdash \psi \,/\, \varphi : M \qquad \Delta \vdash \varphi : N}{\Gamma, \Delta \vdash \psi : MN}\,[/E]$$

- semantyka — język reprezentacji znaczenia;

$$\text{całą świnię} \vdash \mathrm{acc} : \overset{1}{\underset{s:\text{świnia}}{\odot}}\ \mathrm{cały}(s)$$

- twierdzenie o poprawności i pełności systemu z koordynacją i spójnikami addytywnymi.

Dziękuję za uwagę!