Dũng Vũ

Tiếng Việt và vấn đề dịch máy (2)

Các phân tích phần trước đã chỉ ra hai vấn đề nổi bật của máy dịch từ tiếng Việt sang ngoại ngữ: ngữ nghĩa và cấu trúc, cụ thể là *ngữ nghĩa bất định* (nhập nhằng) và *cấu trúc câu sai*. Xin lưu ý, ở đây mới xét vài ví dụ thử nghiệm và trường hợp dịch một chiều từ Việt sang Anh.

3. Vấn đề ngữ nghĩa

Ngữ nghĩa bất định thường xảy ra khi dịch từng từ riêng lẻ. Có hai lý do, một là bản thân từ chứa nhiều nghĩa; hai là máy thiếu khả năng nhận diện.

Ngữ nghĩa từ có thể được nhận diện chính xác hơn nhờ nhiều thông tin: ***từ điển*** và ***văn cảnh*** (cấu trúc câu, cách hành văn theo tình huống, …).

3.1. Thông tin từ điển

Đối với dịch máy, công việc cần thiết đầu tiên là xây dựng một *kho dữ liệu từ vựng (lexical database)* có phẩm chất tốt [1]. Có nhiều cách thực hiện tùy ứng dụng. Một trong những cách là chia nhỏ. Chẳng hạn, mỗi ngành chuyên môn có từ điển riêng, từ thông dụng có từ điển riêng, … Lợi điểm của cách này là số lượng từ có giới hạn, dễ kiểm soát, truy cập nhanh. Ý bài viết muốn đề cập đến từ điển thông dụng.

Xét một ví dụ. Giả sử muốn cho người ngoại quốc hiểu thấu chữ "*anh*" của tiếng Việt, chúng ta sẽ giải thích thế nào ? Giải thích cho người hiểu thế nào thì sẽ dạy cho máy giống vậy (theo phương pháp *trí thông minh nhân tạo*).

"anh" có nhiều nghĩa:

- Là tiếng xưng hô (*nhân vật đại danh từ (personal pronoun)*), ngôi 1, 2, 3, số ít, chỉ phái nam, chứa những nét (tinh thần) ngữ nghĩa đặc trưng:

  - *cực tính tốt (positive polatity)*
  - thân mật trong quan hệ tình yêu nam nữ, vợ chồng, gia đình (anh lớn), trong giao tiếp ngoài xã hội (đối với người lớn hơn, không quá cách biệt tuổi tác; hoặc lịch sự, dùng cho người trưởng thành, có tính ngang hàng).

  Nếu dùng cho ngôi 3, *"anh"* được hiểu là *"anh ta", "anh ấy"*..
  Nếu dùng như *sinh cánh (genitive)*, *"anh"* được hiểu là *"của anh".*

- *"Anh"* (viết hoa) là nước rộng lớn và đông dân nhất trong Liên hiệp Vương quốc Anh và Bắc Ireland, nằm ở phía Tây Bắc Âu châu, có thủ đô là Luân Đôn (London). *"Anh"* là *danh từ riêng (proper noun)*, ngôi 3, số ít.

V.v.

3.2. Thông tin văn cảnh

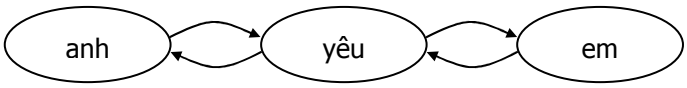
Bản thân từ *"anh"* có nhiều nghĩa, nhưng làm sao để máy hiểu người nói muốn dùng *"anh"* theo nghĩa nào ? Ví dụ câu *"Anh có yêu em không ?"*.

Dựa vào cách làm của trí thông minh nhân tạo, ta có thể giải thích quá trình con người nhận diện ngôn ngữ cho máy hiểu đại để như sau.

Thoạt tiên khi nhận được tiếng *"anh"*, trong tích tắc, người nghe sẽ liên tưởng đến hai trường hợp: "Anh" là nước Anh, hoặc "anh" là tiếng xưng hô.

Nhận được tiếng *"yêu"*, người nghe biết rõ hơn, *"anh"* không phải là nước Anh mà tiếng xưng hô thân mật; *"anh"* là chủ ngữ (chủ từ), nhờ đó có thể tiên đoán bổ ngữ (túc từ) có lẽ là một người nữ. Nhận được tiếng *"em"* thì xác định được đó là người nữ.

Sự gắn bó của ba tiếng *"anh", "yêu", "em"* có thể được cất giữ như một *vốn ngữ liệu (corpus)*) [2] với mạng ngữ nghĩa:



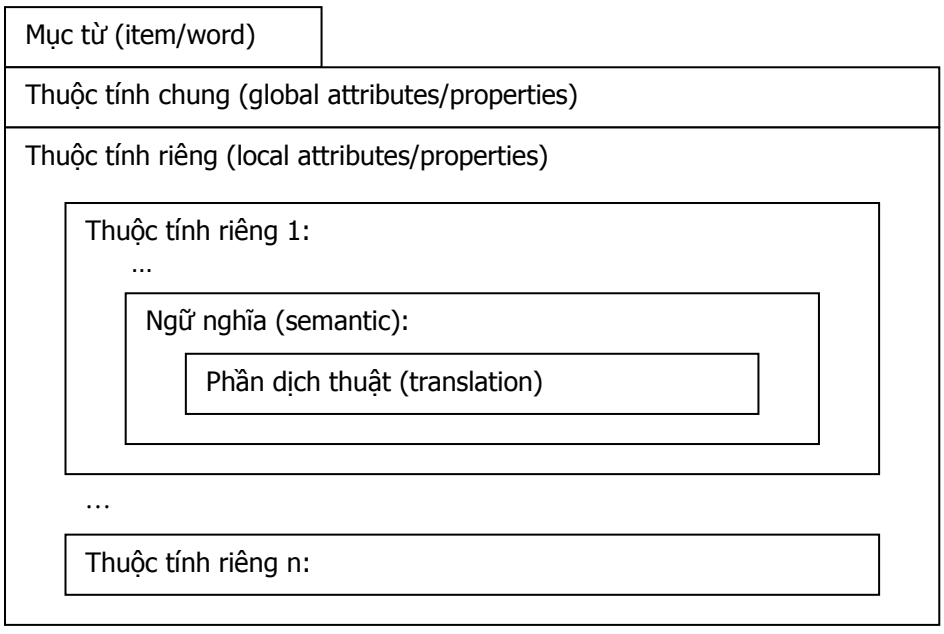
Nhờ mạng ngữ nghĩa này, máy học được cách hiểu như người (tức quá trình *tiếp thụ kiến thức (knowledge acquisition)*) và dịch tự nhiên như người.

3.3 Lưu trữ thông tin

Sau khi đã nắm bắt thông tin của từ và thông tin liên hệ (nhiều như có thể), bước kế tiếp là lưu trữ. Nhiều câu hỏi được đặt ra:

- Dưới dạng thức nào (INI-format, XML-format, SQL-tables, …) ?
- Ở đâu (ASCII-files, XML-files, database, v.d. Oracle, …) ?
- Lập trình bằng ngôn ngữ nào ? (Java, C++, C#, LISP, …) [3]
- Hệ thống hóa ra sao ?
- Cần những quy chiếu (reference), ghi chú, … gì ?
- Tối ưu hóa: cách truy cập dữ liệu, cách xếp đặt *thuộc tính (attributes)* dữ kiện để tránh sự trùng lặp, …
- Muốn mở rộng phải làm sao ? Bài bản *kiến trúc mở (open architecture)* thế nào ?
- *Giao điểm (interface)* cho ngôn ngữ khác (chẳng hạn trong phạm trù tự vựng *Noun*, tiếng Việt có từ loại mà tiếng khác không có, v.d. *danh từ hình thức (formal noun)*, … ; phải đặt tên và sắp xếp ra sao cho ăn khớp với giao điểm chung ?)
- Có khả năng *lập đồ (configure)* không ? V.v.

Có thể mường tượng một mô hình kho dữ liệu từ vựng tổng quát như sau:



Ví dụ lưu trữ mục từ *"anh"*:

anh

```
GLOBAL:
   CAT: noun
   NUM: singular
   SEX: masculine
   CASE: none
   POL: positive
   REG: ba miền (Bắc, Trung, Nam)
   PRONUNC: → anh.wav
LOCAL(1):
   SUBCAT: personal pronoun
   PERS: 1, 2, 3
   SEM: Tiếng xưng hô
      APPL(1): thân mật, dùng trong quan hệ nam nữ, vợ chồng; người nam
      tự xưng anh; người nữ gọi lại cũng bằng anh.
         EXP(1): Anh yêu em
         MORE → em, mình, vợ, chồng, vợ chồng
      APPL(2): thân mật, cũng dùng trong gia đình, họ hàng, xã hội; người
      nam tự xưng anh; người trong vai về em cũng gọi lại bằng anh.
         MORE → em, chị
      APPL(3): lịch sự, ngang hàng, dùng trong giao tiếp xã hội để gọi
      một người nam đối diện từ tuổi trưởng thành trở lên.
         EXP(1): Chào anh Hùng. Tôi tên là Cường. Hân hạnh được biết anh.
         MORE → tôi
      APPL(4): lịch sự, dùng chỉ một người nam ngôi thứ 3, xác định được
      nhờ ngữ cảnh phía trước.
         EXP(1): Hùng là giáo sư đại học. Anh hiện sống ở Hoa Kỳ.
         MORE → anh ta, anh ấy, ảnh

      // phần dịch thuật
      TRANS:
         ENG:
            PERS == 1:
               CASE: NOM = "I"; ACC = "me";
            PERS == 2:
               CASE: NOM = "you"; ACC = "you";
            PERS == 3:
               CASE: NOM = "he"; ACC = "him";
         GER:
            PERS == 1:
               CASE: NOM = "ich"; ACC = "me"; DAT = "mir";
            PERS == 2:
               CASE:
                  if (SEM == 3)
                     NOM = "Sie"; ACC = "Sie"; DAT = "ihnen";
                  else
                     NOM = "du"; ACC = "dich"; DAT = "dir";
            PERS == 3:
               CASE: NOM = "er"; ACC = "ihn"; DAT = "ihm";
         …

LOCAL(2):
   PERS: 3
   NUM: singular
   SEM:
      APPL(1): Người con trai trong gia đình lớn tuổi hơn các người con
      khác,
         // phần dịch thuật
         TRANS:
            ENG: "elder brother"
            GER: "älterer Bruder"
      APPL(2):  hoặc trong họ hàng cùng thế hệ có vai vế cao hơn.
```

```
            // phần dịch thuật
            TRANS:
               ENG: "elder cousin"
               GER: "älterer Cousin"
               …
            MORE → anh họ

   LOCAL(3):
      PERS: 3
      SEM: Người, người nam ngôi thứ 3.
         EXP(1): Có người ví von: "Việt Nam, Cuba như là Trời Đất sinh ra.
         Một anh ở phía Đông. Một anh ở phía Tây. Chúng ta thay nhau canh
         giữ hòa bình cho thế giới. Cuba thức thì VN ngủ. Việt Nam gác thì
         Cuba nghỉ." (Nguyễn Minh Triết). AUDIO & VIDEO →
         http://www.youtube.com/watch?v=fJtl4lApWIU

      // phần dịch thuật
      TRANS:
         ENG:
            CASE: NOM = "one", "man";
         GER:
            CASE: NOM = "Man", "Mann"
         …
   LOCAL(4):
      SUBCAT: adjective
      VIETCAT: processive adjective
      PERS: 1, 2
      GEN: true
      SEM: Của anh (sở hữu từ dùng cho một người nam ngôi 1, 2, 3)
         → LOCAL(1) → SEM → APPL(1), APPL(2), APPL(3)

      // phần dịch thuật
      TRANS:
         ENG:
            PERS == 1:
               CASE: NOM = "my";
            PERS == 2:
               CASE: NOM = "your";
            PERS == 3:
               CASE: NOM = "his";
            …
         GER:
            PERS == 1:
               CASE: NOM = " mein";
            PERS == 2:
               CASE: (SEM == 3 ? NOM = "ihr": NOM = "dein";
            PERS == 3:
               CASE: NOM = " sein";
            …
```

Anh

```
   CAT: noun
   SUBCAT: proper pronoun
   NUM: singular
   POL: neutral
   SEM: Anh quốc, nước rộng lớn và đông dân nhất trong Liên hiệp Vương quốc
   Anh và Bắc Ireland, nằm ở phía Tây Bắc Âu châu, có thủ đô là Luân Đôn
   (London).
   TRANS:
      ENG:
         CASE: NOM = "England"; ACC = "England";
      GER:
         CASE: NOM = "England"; ACC = "England"; DAT = "England";
      GEO:
         IMG(1):
```

IMG(2):

MORE(1) → http://www.weltkarte.com/europa/landkarte_england.htm

HIST: (giới thiệu lịch sử nước Anh ở đây)
MORE(1) → http://www.britannien.de/Geschichte/Geschichte.htm

v.v.

Xin lưu ý:

- Cách trình bày bên trên chỉ nhằm minh họa cho dễ mường tượng thông tin, thuộc tính cần thiết của từ. Trên thực tế, cách mã hóa thông tin cô đọng và khó đọc hơn. Cách lưu trữ dữ liệu cũng khác (như đã nói), tùy người thực hiện. Đây là vấn đề kỹ thuật tin học.
- Ký hiệu dùng miêu tả chỉ là *mã giả (pseudo code),* không nhất thiết theo đúng cú pháp một ngôn ngữ lập trình nhất định.
- Để dễ nhìn toàn cảnh, phần dịch thuật TRANS không được trình bày chi tiết. Phần chi tiết sẽ được nói sau.

Chú thích:

GLOBAL: thuộc tính chung
LOCAL(1): 1st local (thuộc tính riêng, trường hợp 1)
CAT: Lexical category (phạm trù từ vựng)
NUM: Number (số: plural (số nhiều), singular (số ít))
SEX: Gender (giới tính: masculine (nam), feminine (nữ), neutral (trung hòa))
CASE: Case (cách: ACC = accusative (đối cách), NOM = nominative (danh cách), DAT = dative (tặng cách))
POL: Polarity (cực tính: positive (tốt), negative (xấu), neutral (trung hòa))
REG: Region (tính địa phương)
PRONUNC: Pronunciation (cách phát âm)
SUBCAT: Lexical subcategory (tiểu phạm trù từ vựng)
VIETCAT: Vietnamese lexical category (phạm trù từ vựng đặc biệt của tiếng Việt)
PERS: Personal (ngôi thứ nhân vật)
GEN: Genitive (sở hữu chủ)
SEM: Semantic (ngữ nghĩa)
APPL(1): 1st application (ứng dụng 1)
APPL(2): 2nd application (ứng dụng 2)
EXP(1): 1st example (thí dụ 1)
EXP(2): 2nd example (thí dụ 2)
AUDIO: Âm thanh
VIDEO: Phim ảnh

GEO: Geometry (địa lý)
IMG(1): 1st image (hình 1)
IMG(2): 2nd image (hình 2)
HIST: History (lịch sử)
MORE: More information (xem chi tiết)
→ : reference, see (quy chiếu, xem) dẫn qua (hyperlink) những thông tin liên hệ.

Mô hình trên giới thiệu tổng quát, một từ có những điểm chung và điểm riêng.

Điểm chung có thể được tổng hợp thành một *lớp căn bản (base class)*, mang tính chung (GLOBAL): CAT, NUM, SEX, CASE, POL, REG, PRONUNC, … Những thuộc tính này, mỗi từ đều cần.

Điểm riêng thuộc về thuộc tính riêng (LOCAL): VIETCAT, APPL(1), APPL(2), …. Nói thế không có nghĩa là bỏ qua được, bởi mỗi từ cần được dịch. Muốn dịch - tức là muốn hai ngôn ngữ bắt tay nhau - thì phải có *giao điểm* ăn khớp nhau.

Trong câu *"Anh yêu em" = "I love you". Đại danh từ "anh"* ăn khớp với *đại danh từ "I"*; *động từ "yêu"* ăn khớp với *động từ "love"* ; *đại danh từ "em"* ăn khớp với *đại danh từ "you"*.

"anh" vốn là một *đại danh từ* nhưng khi hiểu theo nghĩa *"của anh"*, thì đó lại là một *sở hữu từ (processive adjective)* tương đương với *"my", "your", "his"* của tiếng Anh. Nếu chấp nhận đại danh từ là giao điểm, chúng ta phải thay đổi thuộc tính *processive adjective* của VIETCAT thành *pronoun* cho SUBCAT. Nếu lấy CAT làm giao điểm thì không cần, vì CAT đã mang giá trị *noun*. Đây chỉ là vấn đề kỹ thuật nhỏ nhặt.

(còn tiếp)

Chú thích:

[1] Điều này đã được gợi ý trong bài viết *Vấn đề đánh dấu thanh tiếng Việt* (đã đăng trên talawas bộ cũ) http://www.talawas.org/talaDB/suche.php?res=7657&rb=06

[2] Xem một ví dụ ứng dụng (Automatisch berechnete Kollokationen aus dem DWDS Kerncorpus) ở: http://www.dwds.de/?kompakt=1&qu=unterminieren

[3] Trong lĩnh vực xử lý ngôn ngữ tự nhiên, hiện chưa có một ngôn ngữ lập trình tiện lợi cho tiếng Việt, kể cả những nhu phẩm hỗ trợ, như Qt. Ví dụ về chức năng tìm kiếm. Một ngôn ngữ lập trình có thể xác định một ký tự (character), nhưng không xác định được dấu (sắc, hỏi, huyền, ngã, nặng). Hoặc một từ ghép. "Việt Nam" là hai chữ viết rời. Nếu có gạch nối ("Việt-Nam"), máy dễ tìm hơn. QString của Qt là một phương pháp (method) rất phong phú, nhưng vẫn không có khả năng vừa kể. Mọi chức năng như thế cần được lập trình riêng cho tiếng Việt.