Translating implicit elements in RBMT

Irina Burukina
ABBYY
RSUH, Computational Linguistics
department

ABSTRACT

The present paper addresses MT of asymmetrical linguistic markers, in particular zero possessives. English <-> Russian MT was chosen as an example; however, obtained results can be applied to other language pairs (English – German / Spanish/Norwegian etc.).Overt pronouns are required to mark possessive relations in English. On the contrary, in Russian implicit possessives are regularly used, thus making it very important to analyze them properly, not only for MT but also for other NLP tasks such as NER, Fact extraction, etc. However, concerning modern NLP systems the task remains practically unsolved. The paper examines how modern English <-> Russian MT systems process implicit possessives and explores main problems that exist concerning the issue. As no SB approach can process IP constructions properly, linguistic rules need to be developed for their analysis and synthesis;the main properties of IPs are analyzed to that end. Finally, several rules to apply to RB or model-based MT are introduced that help to increase translation accuracy.

The present research is based on ABBYY Compreno © multilanguage NLP technologies that include MT module.

1. Introduction

MT industry has been recently growing and delivering ever better results. However, several crucial problems remain that prevent us from saying *The perfect MT is achieved.* Among them there are inherent linguistic problems: bilingual lexical ambiguity, bilingual structural ambiguities, structural asymmetries etc. [Hutchins 2007]

One of the possible reasons for structural asymmetry is zero elements (zero subjects, determinants, etc.) allowed in one languages and prohibited in others. For example, Spanish allows zero subjects, while in English overt pronouns are needed. Spanish -> English translation thus requires reconstruction of appropriate overt pronouns and English -> Spanish translation should include deletion of explicit elements, mostly at the beginning of the sentences.

1. Marco calentó el agua del té. Ahora tiene miedo de quemarse.

   Marco warmed water for tea. Now he is afraid to burn himself.

The present research examines how modern MT systems can deal with sentences with structural asymmetry. In particular, we focus on possessive markers in English - Russian language pair as one of the least studied problems.

Implicit possessives (IPs) are zero possessive pronouns, used with inalienable nouns (kinship terms and body-part nouns) in the positions that can be occupied by overt possessives (pronominals or reflexives). The high frequency of inalienable nouns (for example, 1200.6 IPM for the word *рука* 'hand', 484.1 IPM for the word *отец* 'father' [Lyashevskaya, Sharoff 2009]) increases frequency of IP constructions; therefore it becomes crucial for MT system efficiency to take them into account.

This research is based on Compreno multilanguage NLP technologies that include but are not limited to model-based MT. Compreno provides opportunities for NER, Fact extraction, ontology creation etc. [Zuev et al. 2013] It turns out that taking structural asymmetry and IP into account is also important for these tasks and in particular for:

- Text analysis and situation modeling (including interpretation of elliptical structures):

    2. Петя позвонил маме, и Маша тоже.

       Peter.NOM called mother.DAT and Masha.NOM too

       'Peter called his (Peter's) mother and Mary called her (Mary's) mother.'

- Anaphora resolution;

- (Co)reference resolution:

    3. Петя позвонил маме. Машина мама всё слышала.

       Peter.NOM called mother.DAT Mary.PossADJ mother.NOM everything heard

    'Peter called his mother {PERSON-1}. Mary's mother {PERSON-2} heard everything.' The problem of IPs remains almost unsolved. To approach the problem, we have analyzed how the most well-known in Russia SB, RB and model-based MT systems process IPs.

2. Automatic translation of implicit possessives in English – Russian

Processing IPs correctly is essential for English <-> Russian MT systems. As Russian IPs are not unique, the present research can also help to improve MT of other language pairs such as Norwegian <-> English, Spanish <-> English, German <-> English, Russian <-> French etc.

The problem of IPs can be broken down into two primary tasks.

First, in English -> Russian translation overt pronouns should be deleted to get more accurate results. IPs in Russian in many contexts are not only allowed but preferred, especially with body-part nouns.

4. Петя сломал ногу.

   Peter broke leg.ACC

   ‘Peter broke his leg.’

   ?Петя сломал свою ногу.

   Peter broke his.SELF leg.ACC

   ‘Peter broke his leg.’

Second, in Russian -> English translation explicit possessives should be synthesized instead of implicit ones considering properties of their antecedents. Anaphora resolution for IPs in Russian is necessary to synthesize appropriate overt pronouns in English.

5. Девочка любит маму.

   Girl loves mother.ACC

   ‘The girl loves her/*his/*their mother.’

We have analyzed the most well-known in Russia English <-> Russian machine translation systems (rule-based as well as statistics-based) and, unfortunately, all of them showed rather poor results analyzing constructions with IP.

6. Петя подошёл к маме.

   Peter.NOM came up to mother.DAT

   ‘Peter came up to his mother.’

   (Google, SBMT) Peter went to my mother.

   (Yandex, SBMT) Petya went to her mother.

   (SYSTRAN, Hybrid MT) Pete approached the mom.

   (PROMT, RBMT) Petya approached to mother.

   (Compreno, Hybrid MT) Petya walked over to the mother.

7. Peter will be happy if Masha talks to her mother.

   (Google) Питер будем рады, если Маша разговаривает с матерью.

   ‘Peter (we) will be glad if Masha talks to mother.’

   (Yandex) Питер будет счастлив, если бы Маша говорит ей мать.

   ‘Peter will be happy if Masha talks to her (non-reflexive) mother.’

   (SYSTRAN) Питер будет счастливо, если Masha говорит к ее матери.

Peter will it be happy if Masha talks towards her mother.'

(PROMT) Питер будет счастлив, если Маша будет говорить со своей матерью.

'Peter will be happy if Masha is talking to her (reflexive) mother.'

(Compreno) Питер будет счастлив, если Маша будет говорить со своей матерью.

'Peter will be happy if Masha is talking to her (reflexive) mother.'

As is evident from these examples, SBMT cannot process IP properly. It leaves no opportunity for anaphora resolution and appropriate overt pronouns deletion. Altogether we tested almost one hundred typical IP examples, and only most frequently used collocations (idioms) were processed correctly by Google and Yandex MT systems.

8. a. Петя засунул руку в коробку.

(Google) Peter put his hand into the box.

(Yandex) Peter put his hand in the box.

b. Петя надел на голову шляпу.

(Google) Peter put on his head a hat.

(Yandex) Petya put the hat on his head.

c. Peter put his hand into the box.

(Google) Петр положил руку в коробку.

'Peter put his hand into box.'

(Yandex) Петр сунул руку в коробку.

'Peter put his hand into box.'

Current research includes several experiments on bilingual English – Russian corpus attempting to automatically extract sentences with IP. Both recall and precision are insufficient; lots of "noise" sentences were found.

As for the modern RB or model-based MT systems, most of them usually contain no rules for IP or lack accuracy. The present paper provides a detailed description of ABBYY Compreno MT system as an example.

3. The Compreno system

ABBYY Compreno MT system is part of ABBYY Compreno multilanguage NLP technologies.

Compreno represents a hybrid model-based MT system. The core of the system is the Universal Semantic Hierarchy (USH), a thesaurus-like structure with universal semantic concepts

as its nodes,that provides semantic analysis of the text. Complete syntactic analysis represents the second part of the technology.

While translating the input text, Compreno analyzes syntactic structures of each sentence that are further converted into intermediate semantic structures. During synthesis semantic structures are converted into output sentences. All the structures created by the system are trees. The crucial idea for Compreno is that while syntactic (surface) structures are language dependent, at the semantic (deep) level nodes of the trees are concepts from the USH, and arcs are universal relations between these concepts. Apart from this, the system establishes non-tree links between nodes of the tree to represent anaphora, coordination, etc.

Compreno is claimed to be a system applicable to any natural language. Language-dependent rules have to be created anew; however, universal part of the system fits different language models.

Below we give examples of syntactic semantic structures for the sentence *The boy showed his file to the girl*. For more information on the project see [Anisimovich et al. 2012, Zuyev et al. 2013].

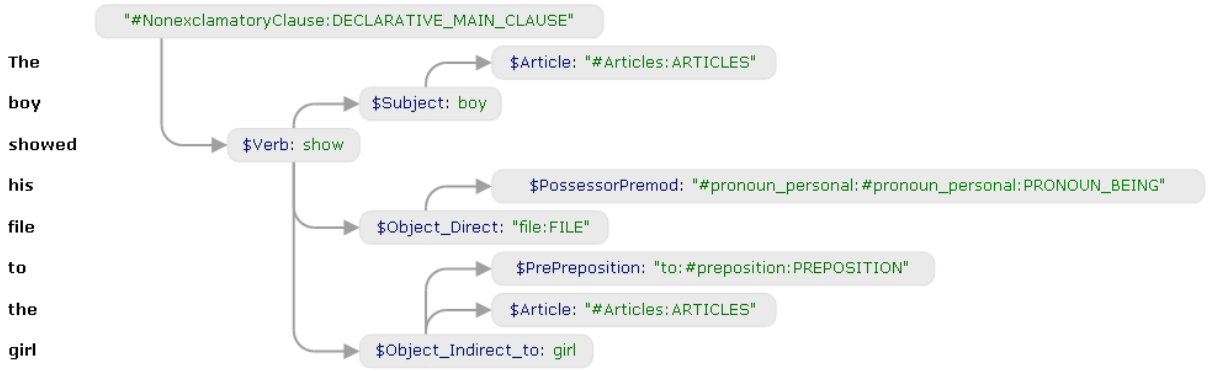

Figure 1: Syntactic structure

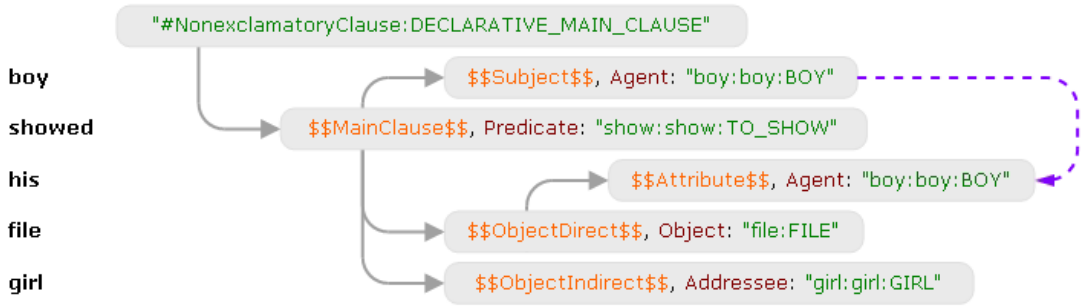

Figure 2: Semantic structure

Normally, special diathesis is used to convert language specific syntactic structures to universal semantic structures. Diathesis is the description of correspondence between elements of two structures, for example, between syntactic positions and semantic slots (syntactic subject = semantic agent etc.). However, sometimes it becomes very difficult or almost impossible to get a semantic structure that can be further converted to the output syntactic structure because of language asymmetries. In these cases at the analysis stage one or several syntactic positions can

be removed, replaced or added to get an appropriate result. This is done with transformation rules.

Transformation rules (TR) apply to a part of a syntactic tree. Usually TR consists of several productions, describing different possible operations on syntactic components. The system tries to implement each of these productions one by one. If attempt is successful the system passes onto the next TR. The productions can change properties of the components, replace them, delete or add new ones.

As for the possessive constructions, TRs mostly describe external possessors. [Bogdanov, Leontyev 2013] For the implicit possessives analysis there is only one TR. Its productions describe several particular contexts in which Russian IP most commonly appear. At the synthesis stage the system allows deletion of overt possessive pronouns only for body-part nouns.

During Russian -> English translation Compreno restores overt possessives in several cases listed below.

I. Body-part noun with IP is a semantic orientational locative (Locative_PartAsOrientation); normally agent is chosen as antecedent, however if the main verb describes movement of an object by someone, the object becomes ' candidate antecedent':

9. a. Мальчик лежал головой к окну. -> The boy was lying his head to the window.

   Lit.: Boy was lying head.INSTR to window.

   b. Мальчик поставил стол ножками вверх. -> The boy set a table with his legs upward.

   Lit.: Boy set table.ACC legs.INSTR upwards.

II. External body-part noun (*нога* 'leg', *зубы* 'teeth', *хвост* 'tail') with IP is a semantic instrument, locative or initial point locative; subject is chosen as antecedent:

10. a.Мальчик расколол орех зубами. -> The boy cracked a nut with his teeth.

    Lit.: Boy cracked nut.ACC teeth.INSTR

    b. Мальчик держал в руках мяч. -> The boy kept a ball in his hands.

    Lit.: Boy kept in hands ball.ACC

III. Body-part noun with IP is a semantic object for verbs like *ломать, трогать* ('break', 'touch') and there is no external possessor; subject is chosen as antecedent:

11. Мальчик сломал ногу. -> The boy broke his leg.

    Lit.: Boy broke leg.ACC

IV. Kinship term with implicit possessive is coordinated with non-kinship term; the last one is chosen as antecedent:

12. Президент с супругой приехали в Москву. -> The president and his spouse arrived to Moscow.

    Lit.: President with wife arrived in Moscow

During English -> Russian translation Compreno deletes overt possessives only for body-part nouns if a body-part noun is a semantic orientational locative, semantic instrument, semantic locative inside the noun group with *with*, semantic locative for possession verbs.

13. a. The boy lies with his head to the window -> Мальчик лежит головой в окно.

    Lit.: The boy lies head.INSTR into window

    b. The boy nodded his head. -> Мальчик кивнул головой.

    Lit.: The boy nodded head.INSTR

    c. She fed the baby with her breast -> Она накормила ребёнка грудью.

    Lit.: She fed baby breast.INSTR

    d. A boy with a ball in his hand. -> Мальчик с мячом в руке.

    Lit.: Boy with ball in hand

    e. She had pills in her hands -> У неё были таблетки в руках.

    Lit.: Of her were pills in hands

Compreno's TRs for IP are insufficient and need to be specified and expanded. To do this we should first provide the profound linguistic study of the main properties of IPs and IPs constructions.

4. Definition and main properties of implicit possessives

As we have said before, implicit possessives (IP) are zero possessive pronouns, used with inalienable nouns (kinship terms and body-part nouns) in the positions that can be occupied by overt possessives (pronominals or reflexives). They can be used as either deictic elements (when they point out entities in non-linguistic 'real' context) or proper anaphors (when they are coreferential with entities mentioned in the same text).

14. a. Мамы нет в доме.

    Mother.GEN not in house.PREP

    'My mother is not in the house.'

    b. Что случилось? Рука болит?

    What happened Hand.NOM aches

    'What happened? Is your hand aching?'

c. Петя позвонил маме.

Peter.NOM called mother.DAT

'Peter called his mother.'

As was mentioned earlier, Russian IPs are not a unique phenomenon. Zero possessive markers appear rather frequently in different world languages, European and non-European alike. Several examples are listed below:

- Implicit possessives with body-part nouns in Norwegian:
  15. Han stakk hendene i lomma.

      He put hands.DEF in pocket.DEF

      'He put his hands in the pocket.' [Loedrup 2010]

- External dative possessors with body-parts nouns in German:
  16. Sie wäscht sich das Gesicht.

      She washes self face.DEF

      'She washes her face.' [Loedrup 2012]

- Zero possessive markers with kinship terms in Dogon:
  17. u ba

      you father

      'your father' [Haspelmath 2008]

Current research studies the behavior of noun phrases (NP) with IPs in different contexts in Russian and introduces the main properties of IPs that are listed below.

- The antecedent of IP (i.e. possessor) should be human;
- IP can point out persons in non-linguistic "real" context. However, they can indicate only the speaker (used in assertive direct speech) or the hearer (used in imperatives or questions) (1a,b);
- Distance between the NP with IP and its potential antecedent in fact doesn't matter much. IP can be locally bound by its antecedent, located in the same minimal clause. It can also be coreferential with element in another clause. A local antecedent is preferable. However, if it's not a subject then the semantically appropriate subject of the matrix clause should be chosen:
  18. Маша хотела, чтобы мама позвонила Пете.

      Mary.NOM wanted that mother.NOM called Peter.DAT.

      'Mary wanted her mother to call Peter.'

- The preferable antecedent is the subject of the clause;

- Used in elliptical contexts or in the sentences with quantifiers, IPs can get sloppy reading only:

19. a. Петя почесал руку, и Маша тоже.

    Peter.NOM scratched hand.ACC and Mary.NOM too

    'Peter scratched his hand and Mary scratched her hand.' – *sloppy reading*

    'Peter scratched his hand and Mary scratched his hand too.' – *strict reading, impossible*

    b. Каждый ученик позвонил маме.

    Every pupil.NOM called mother.NOM

    'Every pupil called his own mother.' – *sloppy reading*

    'All pupils called their mother.' – *strict reading, impossible*

Now we can proceed to development of analysis / synthesis rules.

5. Development of analysis/synthesis rules for implicit possessives

5.1. How general can these rules be?

The initial idea was to propose a small number of the most general rules. However, it proved infeasible. There are several problems that hold us back from analyzing every sentence unambiguously without exceptions. Large-scale MT systems like Compreno cannot risk the precision of translation even to obtain higher recall, more literary results and economical set of TRs.

First, it is impossible to formalize pragmatics. Implicit possessives are essentially ambiguous. In many cases (especially in direct speech) they may be used in deictic function to denote the speaker himself.

20. Вчера студенты позвонили матери.

    Yesterday students called mother.DAT

    'Yesterday students called my mother.'

Second, even though we have analyzed IPs in different contexts and discovered their main properties, these properties should be considered regularities, but not strict rules. There are always exceptions that can be crucial for the evaluation.

21. a. Маша попросила Петю позвонить маме.

    Masha asked Peter to call mother.DAT

    'Masha asked Peter to call her mother.'

    b. Учительница попросила Петю позвонить маме.

Teacher asked Peter to call mother.DAT

'The teacher asked Peter to call his (Peter's) mother.'

Third, semantics of predicates should also be taken into account.

5.2. What rules can be proposed to improve translation

Nevertheless, we propose several rules that help improve the translation.

The present paper shows several rules that should be used during analysis stage to improve the results of Russian -> English MT. What is even more important is that these rules can be applied to the Russian IP analysis in general.

- In direct speech first person possessive pronoun should be inserted with kinship terms and body-part nouns; in imperative sentences and questions second person possessive pronound should be inserted;
- If a body-part noun with IP is the subject of subordinate clause, the subject of the main clause should be considered as preferable antecedent if there is no external possessor construction), the appropriate overt pronoun should be inserted;
- If a kinship term with IP plays the ContrAgent semantic role of "social interaction" predicate, the subject of that predicate should be considered preferable antecedent, and the appropriate overt pronoun should be inserted:
  22. Петя спорил с мамой.

      Peter.NOM argued with mother.INSTR

      'Peter was arguing with his mother.'

As for English -> Russian MT we discovered a phenomenon that allows us to propose a more general rule. In several cases the deletion of pronoun helps not only to get more literary results, but also to preserve the initial sense of sentences. We argue that this approach can be applied as well to MT of other language pairs. In English there are no possessive reflexives and only pronouns like *his, her* are used. The sentences with such pronouns can be ambiguous.

23. Bill asked John to call his father.

However, while translating such sentences to Russian, the system should unambiguously identify the antecedent; either reflexive *svoi* or a pronominal is used. One of interpretations is missed and incorrect translation can be got.

The rule proposed below should be used at the analysis stage during English -> Russian MT.

In simple sentences if the subject (A) and the direct object (B) share the same grammatical and semantic characteristics, the kinship term with overt possessive is used in the same sentence and A or B is chosen as antecedent to that possessive, the pronoun should be deleted.

In composite sentences if the subject of the main clause (A) and the subject or the direct object of the subordinate clause (B) share the same grammatical and semantic characteristics,

the kinship term with overt possessive is used in subordinate clause and A or B is chosen as antecedent to that possessive, the pronoun should be deleted.

24. a. Bill introduced John to his father.

    b. Mary asked Jane to call her mother.

    c. The boys wanted the girls to show the books to their mothers.

The rule has been successfully applied to ABBYY Compreno MT system and verified.

6. Conclusion

This research is focused on formerly unsolved problem of English <-> Russian MT of implicit possessives however the proposed rules can be applied to other language pairs and can help as well to solve text analysis tasks in general.

The research is carried out on the basis of ABBYY Compreno NLP technologies and the obtained results allow to increase MT system efficiency.

As it turns out to be almost impossible to process IP statistically we had to start with theoretical linguistic research. Investigating main properties of IPs in Russian we managed to formulate several rules for automatic analysis and synthesis of IP constructions. Among them the pronoun deletion rule that helps to get more precise translation of English sentences into Russian or another language allowing zero possessives.

A more detailed study of the IP phenomenon is being carried out. Below we describe the main tasks of our work in progress:

- Investigation of IPs in other languages (Norwegian, German, Spanish, etc.); verification of the “deletion” rule to other language pairs;
- Implementation of proposed rules for other but MT NLP tasks;
- Further analysis of Russian IP including examination of semantics of predicates in sentences with IP.

Acknowledgements

We are particularly grateful to Vladimir Selegey (ABBYY) and Alexey Leontyev (ABBYY) for their help and support. A special thank goes to prof. Yakov Testelets (RSUH) for his help and advices on theoretical part of research.

References

Anisimovich, K. V., K. Ju. Druzhkin, F. R. Minlos, M. A. Petrova, V. P. Selegey and K. A. Zuev. 2012. “Syntactic and semantic parser based on ABBYY Compreno linguistic technologies.” *Computational linguistics and intellectual technologies*, *11*. Moscow: RSUH.

Bogdanov, A. V. and A. P. Leontyev. 2013. “Description of the Russian external possessor construction in a natural language processing system.” *Computational linguistics and intellectual technologies*, *12*. Moscow: RSUH.

Haspelmath, M. 2008. *Alienable vs. inalienable possessive constructions*. Presentation at Leipzig Spring School on Linguistic Diversity.

Hutchins, J. 2007. “Machine translation in Europe and North America: current state and future prospects.” *JAPIO 2007 Yearbook.* Tokyo: Japan Patent Information Organization.

Lødrup, H. 2010. “Implicit possessives and reflexive binding in Norwegian.” *Transactions of the Philological Society,* 108:2.

Lødrup, H. 2012. *Inalienables in Norwegian and binding theory.*

ttp://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.7636

Lyashevskaya, O. N. and S. A. Sharoff. 2009. *Russian frequency dictionary*. Moscow: Azbukovnik.

Zuev K. A., E. M. Indenbom and M. V. Yudina. 2013. “Statistical machine translation with linguistic language model.” *Computational Linguistics and Intellectual Technologies*, *13*. Moscow: RSUH.